Федеральное агентство по образованию
Уральский государственный технический университет – УПИ
имени первого Президента России Б.Н. Ельцина

С.М. БОРОДАЧЁВ

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Учебное пособие

Научный редактор – проф., д-р физ.-мат. наук О.И. Никонов

Печатается по решению редакционно-издательского совета
УГТУ – УПИ от 26.05.2009 г.

Екатеринбург
УГТУ – УПИ
2009

УДК 519.237(075.8)
ББК 22.172я73
Б83

Рецензенты:
кафедра информационных систем в экономике УрГЭУ (зав. кафедрой – проф., д-р физ.-мат. наук А.Ф. Шориков);
директор негосударственного образовательного учреждения "Новые технологии и программы", д-р физ.-мат. наук В.И. Зенков.

Бородачёв, С.М.
Б83 Многомерные статистические методы: учебное пособие / С.М. Бородачёв. Екатеринбург: УГТУ – УПИ, 2009. 84 с.

ISBN 978-5-321-01613-8

Пособие содержит теоретический материал по многомерным статистическим методам, упражнения по всем разделам курса и лабораторный практикум, подводящий студентов к решению реальных задач исследования зависимости разнообразных показателей, распознавания, классификации.

Предназначено для студентов информационно-математических и экономических специальностей.

Библиогр.: 12 назв.

ISBN 978-5-321-01613-8 УДК 519.237(075.8)
ББК 22.172я73

Оглавление

Многомерные статистические методы – раздел математической статистики, содержащий методы анализа случайного вектора по его выборке. При этом множественный регрессионный анализ обычно изучают в эконометрике.

1. Корреляционный анализ

Корреляционный анализ - метод, позволяющий обнаружить зависимость между несколькими случайными величинами.

1.1. Корреляционный анализ количественных переменных

Пример 1.1: путём опроса 10 студентов были собраны сведения:

n	Оценка на экзамене, x_1^n	Число пропусков занятий, x_2^n	Число двоек в семестре, x_3^n
1	5	1	3
2	2	8	2
6	3	6	4
4	2	12	1
5	3	9	1
6	4	2	5
7	4	0	6
8	5	0	4
9	4	3	4
10	2	15	2

Ставится задача: по этим данным установить, есть ли зависимость между случайными величинами X_1 – оценка на экзамене, X_2 – число пропусков занятий, X_3 – число двоек в семестре для случайно выбранного (любого) студента.

Введём случайный вектор

$\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$. Тогда $\vec{x}^1 = \begin{pmatrix} 5 \\ 1 \\ 3 \end{pmatrix}, \ldots, \vec{x}^n, \ldots, \vec{x}^N$ – полученные в наблюдениях значения вектора $\vec{X}$. $N = 10$.

Определение:

$\operatorname{cov}(X_i, X_j) = M(X_i - MX_i)(X_j - MX_j) = MX_iX_j - MX_iMX_j$

– ковариация случайных величин X_i и X_j.

$$\rho(X_i, X_j) = \frac{\operatorname{cov}(X_i, X_j)}{\sqrt{DX_iDX_j}} \qquad (1)$$

– коэффициент корреляции случайных величин X_i и X_j.

$K_{\vec{X}} = (\operatorname{cov}(X_i, X_j))$ – ковариационная и $R_{\vec{X}} = (\rho(X_i, X_j))$ корреляционная матрицы случайного вектора $\vec{X}$.

Теорема:

X_i, X_j – независимы $\Rightarrow \rho(X_i, X_j) = 0$,

X_i, X_j – зависимы $\Leftarrow \rho(X_i, X_j) \neq 0$.

Если бы мы могли установить, что $\rho(X_1, X_2) \neq 0$, то доказали бы зависимость оценки на экзамене и числа пропусков. Но совместное распределение случайных величин X_1, X_2, X_3, нужное для вычисления (1), нам неизвестно, и встаёт задача *оценки* коэффициента корреляции по экспериментальным данным.

Теорема: несмещённой и состоятельной оценкой ковариации случайных величин является статистика

$$\widehat{\operatorname{cov}}(X_i, X_j) = \frac{1}{N-1} \sum_{n=1}^{N} (x_i^n - \overline{x}_i)(x_j^n - \overline{x}_j), \qquad (2)$$

где

$$\overline{x}_i = \frac{1}{N}\sum_{n=1}^{N} x_i^n .$$

Значит, оценка ковариационной матрицы имеет вид:

$$\widehat{K}_{\vec{X}} = \frac{1}{N-1}\sum_{n=1}^{N}(\vec{x}^n - \overline{\vec{x}})(\vec{x}^n - \overline{\vec{x}})^T =$$

$$= \frac{1}{N-1}\left[\sum_{n=1}^{N}\vec{x}^n\vec{x}^{nT} - N\overline{\vec{x}}\,\overline{\vec{x}}^T\right].$$

Если ввести матрицу объект-свойство:

$$z_{N\times k} = \begin{pmatrix} \vec{x}^{1^T} \\ \vec{x}^{2^T} \\ ..... \\ \vec{x}^{N^T} \end{pmatrix} = \begin{pmatrix} x_1^{\,1} & x_2^{\,1} & ... & x_k^{\,1} \\ x_1^{\,2} & x_2^{\,2} & ... & x_k^{\,2} \\ ..... & & & \\ x_1^{\,N} & x_2^{\,N} & ... & x_k^{\,N} \end{pmatrix}$$

и вектор $\vec{1} = (1...1)^T$, то

$$\widehat{K}_{\vec{X}} = \frac{1}{N-1}\left[z^T z - \frac{1}{N} z^T \vec{1}\vec{1}^T z\right]. \qquad (3)$$

Пример 1.1 (продолжение):

$$\widehat{R}_{\vec{x}} = \begin{pmatrix} 1 & -0,904 & 0,629 \\ -0,904 & 1 & -0,789 \\ 0,629 & -0,789 & 1 \end{pmatrix},$$

$\widehat{\rho}_{12} = \widehat{\rho}(X_1, X_2) = -0,904 < 0$. Отрицательный знак, характерный для обратной зависимости, кажется правдоподобным для оценки на экзамене и числа пропусков. Поскольку $/\widehat{\rho}/ \le 1$ всегда, то у $\widehat{\rho}_{12}$ большая абсолютная величина. Что, уже доказана зависимость? Нет, это ведь *оценка* не равна нулю! Необходимо проверить гипотезу об истинном коэффициенте корреляции.

$$H_0 : \rho_{12} = 0$$

$$H_1 : \rho_{12} \neq 0 .$$

$$SL(\text{данных против } H_0) = P\left\{\text{получить наши данные}\middle| H_0\right\} =$$

$$= 2\left[1 - \Phi_{N-2}^{Cm}\left(\frac{|\widehat{\rho}|\sqrt{N-2}}{\sqrt{1-\widehat{\rho}^2}}\right)\right].$$

Эта точная формула предполагает нормальное совместное распределение компонент вектора $\vec{X}$.

$SL = 0{,}00032$

Вывод: *SL* мал (<0,01), данные имеют высокую значимость против H_0. H_0 отвергаем, теперь зависимость можно считать доказанной.

$\hat{\rho}_{13} = 0{,}629$. Положительный знак выглядит странно! Может быть это связано с тем, что величины X_1 и X_3 сильно зависят от третьей величины X_2? Чтобы найти коэффициент корреляции между X_1 и X_3, "очищенный" от влияния X_2, заменим X_1 и X_3 на

$$\tilde{X}_1 = X_1 - M(X_1 \mid X_2), \qquad (4)$$

так как именно функция регрессии является наилучшим предиктором X_1 по X_2, то есть отражает зависимость X_1 от X_2, и если её вычесть, то $\tilde{X}_1$ не будет коррелирована с X_2:

$$\begin{aligned}\operatorname{cov}(\tilde{X}_1, X_2) &= M\tilde{X}_1 X_2 - M\tilde{X}_1 MX_2 = \\ &= MX_1X_2 - M\left[X_2 M(X_1 \mid X_2)\right] - \\ &- \left[MX_1 - M(M(X_1 \mid X_2))\right] MX_2 = 0.\end{aligned}$$

Аналогично $\tilde{X}_3 = X_3 - M(X_3 \mid X_2)$.

Определение: частным коэффициентом корреляции случайных величин X_i, X_j при исключении влияния случайной величины X_k называется величина, определяемая формулой

$$\rho_{ij(k)} = \frac{\operatorname{cov}(\tilde{X}_i, \tilde{X}_j)}{\sqrt{D\tilde{X}_i D\tilde{X}_j}}.$$

Если функции регрессии в формуле (4) считать линейными, т. е.

$$M\left(X_1 \mid X_2\right) = \rho_{12} \frac{\sigma_{X_1}}{\sigma_{X_2}} \left(X_2 - MX_2\right) + MX_1,$$

то получим

$$\rho_{ij(k)} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{1-\rho_{ik}{}^2}\sqrt{1-\rho_{jk}{}^2}}.$$

Доказательство: $D\tilde{X}_1 = M\left(X_1 - \rho_{12}\frac{\sigma_{X_1}}{\sigma_{X_2}}\left(X_2 - MX_2\right) - MX_1 \right)^2 =$

$$\sigma^2_{X_1} - 2\rho_{12}\frac{\sigma_{X_1}}{\sigma_{X_2}}\operatorname{cov}\left(X_1, X_2\right) + \rho^2_{12}\frac{\sigma^2_{X_1}}{\sigma^2_{X_2}}\sigma^2_{X_2} = \sigma^2_{X_1}\left(1-\rho^2_{12}\right).$$

Аналогично преобразуется числитель.

Замечание: частный коэффициент корреляции для произвольного числа всех остальных исключённых влияний:

$\rho_{12(3\ldots r)} = -\frac{A^R_{12}}{\sqrt{A^R_{11}\cdot A^R_{22}}}$, где A^R_{ij} – алгебраическое дополнение элемента ρ_{ij} в определителе $\det\left(R_{\vec{X}}\right)$. Алгебраическое дополнение проще выразить через $\det\left(R_{\vec{X}}\right)$ и элемент обратной матрицы.

Проверка гипотез о частном коэффициенте корреляции:

H_0: $\rho_{12(3\ldots r)} = 0$

H_1: $\rho_{12(3\ldots r)} \neq 0$,

$$SL = 2\left[1 - \Phi^{Cm}_{N-r}\left(\frac{\left|\hat{\rho}_{12(3\ldots r)}\right|\sqrt{N-r}}{\sqrt{1-\hat{\rho}_{12(3\ldots r)}{}^2}}\right)\right].$$

Пример 1.1 (продолжение):

$\hat{\rho}_{13(2)} = -0,326.$ Знак изменился на более правдоподобный, правда,

$SL = 2\left[1 - \Phi^{Cm}_7\left(\frac{0,326\sqrt{7}}{\sqrt{1-0,326^2}}\right)\right] = 0,392$ – велик, частный коэффициент корреляции может быть равен нулю.

Пример 1.2:

$$z = \begin{pmatrix} \text{число} & \text{число} \\ \text{детей} & \text{комнат} \\ \text{в семье} & \text{в квартире} \\ 2 & 2 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}, \quad \vec{X} = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}.$$

Необходимо найти оценку корреляционной и ковариационной матрицы.

$$z^T = \begin{pmatrix} 2 & 0 & 1 \\ 2 & 1 & 1 \end{pmatrix} \quad \vec{1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix},$$

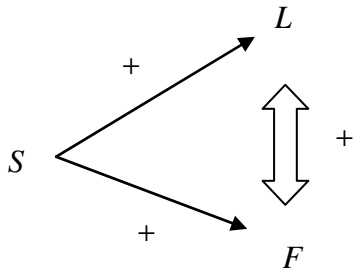
$$z^T z = \begin{pmatrix} 5 & 5 \\ 5 & 6 \end{pmatrix} \qquad z^T \vec{1}\vec{1}^T z = \begin{pmatrix} 9 & 12 \\ 12 & 16 \end{pmatrix},$$

$\widehat{K}_{\vec{X}} = \frac{1}{2}\left[\begin{pmatrix} 5 & 5 \\ 5 & 6 \end{pmatrix} - \frac{1}{3}\begin{pmatrix} 9 & 12 \\ 12 & 16 \end{pmatrix}\right] = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1/3 \end{pmatrix}$ – оценка ковариационной матрицы.

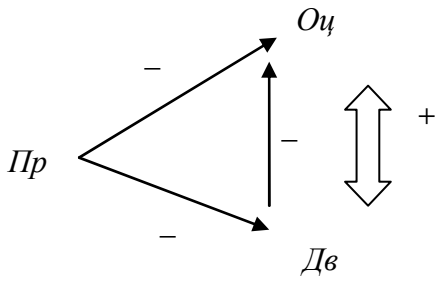
$\widehat{R}_{\vec{X}} = \begin{pmatrix} 1 & \sqrt{3}/2 \\ \sqrt{3}/2 & 1 \end{pmatrix} \approx \begin{pmatrix} 1 & 0{,}87 \\ 0{,}87 & 1 \end{pmatrix}$ – оценка корреляционной матрицы.

Определение: корреляция, которая есть следствие влияния одной или более других переменных, называется ложной корреляцией.

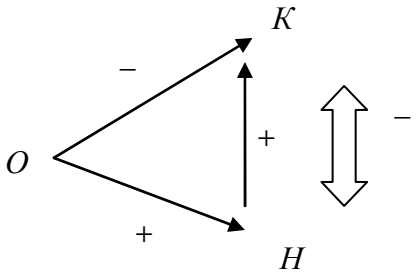
Пример 1.3: пусть L – число жертв пожара, F – число пожарных, участвующих в тушении пожара. Если собрать данные, то $\hat{\rho}(L,F)$ будет больше нуля и значима! Значит ли это, что чем больше послать пожарных на тушение, тем больше будет жертв? Ясно, что L и F зависят от третьей величины S – размера пожара, что вызывает между ними положительную ложную корреляцию:



Пример 1.1 (продолжение): оценка на экзамене и число двоек в семестре находятся в ложной положительной корреляции.



Иногда ложная корреляция может замаскировать (свести к 0) имеющуюся корреляцию: *О* – число обрывов нити ткацкого станка, *К* – качество ткани, *Н* – число наладок станка. *К* и *Н* могут показаться не коррелированными:



Множественный коэффициент корреляции

Пусть целью исследования является мера зависимости одной величины (обозначим её X_1) от всех остальных вместе $X_2,...,X_r$.

Определение

Множественным (сводным) коэффициентом корреляции величины X_1 с остальными $X_2,...,X_r$ называется величина

$$\rho^2_{1(.)} = \frac{\vec{a}^T K^{-1}_{(.)} \vec{a}}{\sigma^2_1}\ , \qquad (4.1)$$

где $a_k = \mathrm{cov}(X_1, X_k), k = 2, ..., r$, $K_{(.)}$ – ковариационная матрица величин $X_2,...,X_r$.

Эквивалентное выражение через полную корреляционную матрицу вектора $\vec{X}$ $R_{\vec{X}}$:

$$\rho^2_{1(.)} = 1 - \frac{1}{[(R_{\vec{X}})^{-1}]_{11}}. \qquad (4.2)$$

Если в этом выражении заменить корреляционную матрицу её оценкой, получим оценку множественного коэффициента корреляции.

Пример 1.1 (продолжение)

Оценка множественного коэффициента корреляции оценки на экзамене с числом пропусков занятий и числом двоек в семестре.

$$[\widehat{R}_{\vec{x}}]^{-1} = \begin{pmatrix} 6.151 & 6.663 & 1.393 \\ 6.663 & 9.871 & 3.603 \\ 1.393 & 3.603 & 2.969 \end{pmatrix},$$

$$\hat{\rho}^2_{1(.)} = 1 - \frac{1}{6.151} = 0.8374.$$

Оценка множественного коэффициента корреляции совпадает с коэффициентом детерминации R^2 в регрессионном анализе КЛММР, если X_1 рассматривать как зависимую переменную, а $X_2,...,X_r$ как независимые переменные (факторы):

$$\hat{\rho}^2_{1(.)} = R^2. \qquad (4.3)$$

О попарной независимости всех компонент случайного вектора (критерий Уилкса – Бартлетта)

H_0: $\vec{X} \sim N(\vec{a}, K)$, $R_{\vec{X}} = I$ – единичная матрица (признаки попарно не коррелированы).

H_1: Признаки попарно коррелированы.

Статистика критерия: $\Gamma = -\left(N - \frac{2r+11}{6} \right) \ln\left(\det\left(\widehat{R}_{\vec{X}} \right) \right)$.

$$SL(\gamma) = P\left\{ \Gamma \geq \gamma \middle| H_0 \right\} = \left| \begin{array}{l} при\ H_0 \\ \mathrm{N} \to \infty \\ \Gamma \sim \chi^2_{\frac{r(r-1)}{2}} \end{array} \right| \approx P\left\{ \chi^2_{\frac{r(r-1)}{2}} \geq \gamma \middle| H_0 \right\} =$$

$$= 1 - \Phi^{\chi^2}_{\frac{r(r-1)}{2}}(\gamma).$$

1.2. Корреляционный анализ ординальных (порядковых) переменных

Ранговая корреляция

Некоторые показатели трудно измерить в каких-то единицах, например, X – целеустремлённость работника. Единственное, что можно сделать, так это ранжировать, т.е. выстроить в порядке возрастания показателя X все N наблюдений. Пусть r_X^n – ранг (место) n-го объекта при ранжировании по показателю X.

Аналогично производится ранжирование по показателю Y, и если между переменными X и Y имеется прямая зависимость, то r_X^n и r_Y^n должны быть близкими (оба малыми или оба большими).

Определение: коэффициент ранговой корреляции Спирмена –

$$\hat{\rho}_S = 1 - \frac{6\sum_{n=1}^{N}\left(r_X^n - r_Y^n\right)^2}{N\left(N^2-1\right)} \quad - \qquad (5)$$

играет ту же роль, что и коэффициент корреляции Пирсона $\hat{\rho}(X,Y)$.

Пример 1.4: при приеме на работу новички проходили письменный тест и собеседование (остались заметки, позволившие проранжировать новичков по итогам собеседования). Через год работы новички были еще раз проранжированы по трудовым успехам.

Результаты ранжирования новичков.

Показатель	Новичок							
	И	С	П	А	З	В	Р	К
Трудовые успехи	1	2	3	4	5	6	7	8
Письменный тест	6	2	7	4	1	5	3	8
Собеседование	1	4	2	3	6	5	8	7

На следующий год решили для отбора новичков оставить что-то одно. Ясно, следует оставить то, что имеет большую корреляцию с трудовыми успехами.

$$\hat{\rho}_s\left(T\ y, П\ m\right) = 1 - \frac{6}{8\left(64-1\right)}\left(25+0+16+0+16+1+16+0\right) = 0{,}119\,.$$

$\hat{\rho}_s\left(T\ y, C\right) = 0{,}881$, $SL = 0{,}0039$. Оставляем собеседование.

Уровень значимости данных можно приближённо находить по той же формуле, что и для коэффициента корреляции Пирсона (хорошо применима при N > 10).

Коэффициент конкордации (согласованности) нескольких порядковых переменных

Пример 1.5: пусть *K* экспертов проранжировали *N* объектов каждый по-отдельности. r_k^n – ранг *n*-го объекта у *k*-го эксперта.

Результаты ранжирования.

	r_1	…	r_K
1	r_1^1		r_K^1
…			
N	r_1^N		r_K^N

Возникает задача проверки совокупной согласованности этих *K* ранжировок (переменных), чтобы при ее наличии выявить некоторое единое упорядочение, или задача выявления некомпетентного эксперта по степени несвязности его ранжировки с остальными.

Определение: М. Кендалл предложил оценивать конкордацию *K* ранжировок коэффициентом конкордации

$$\hat{w}\left(K\right) = \frac{12}{K^2\left(N^3-N\right)}\sum_{n=1}^{N}\left(\sum_{k=1}^{K} r_k^n - \frac{K\left(N+1\right)}{2}\right)^2. \qquad (6)$$

Свойства коэффициента конкордации:

1. $0 \le \widehat{w} \le 1$.

2. $\widehat{w} = 1 \Leftrightarrow$все K ранжировок совпадают.

Проверка гипотезы об истинном коэффициенте конкордации:

H_0: $w(K) = 0$, т. е. нет никакой согласованности,

H_1: $w(K) > 0$.

$$SL = P\left\{\widehat{W}(K) \ge \widehat{w}(K) \middle| H_0\right\} = \left|\begin{array}{l}\text{при } H_0,\ N > 7 \\ K(N-1)\widehat{W}(K) \approx\sim \chi^2_{N-1}\end{array}\right| \approx$$

$$\approx 1 - \Phi^{\chi^2}_{N-1}\left(K(N-1)\widehat{w}(K)\right).$$

Точное распределение статистики $\widehat{W}(K)$ – [6]:

Распределение коэффициента конкордации Кендалла

$$P\{ K^2(N^3 - N)\,\widehat{W}(K)\,/12 \geq s \} = Q_w$$

N=3													
K=3		K=5		K=6		K=7		K=8		K=9		K=10	
s	Q_w	s	Q_w	s	Q_w	s	Q_w	s	Q_w	s	Q_w	s	Q_w
6	0,528	14	0,367	18	0,252	24	0,237	26	0,236	32	0,187	32	0,222
8	361	18	182	24	184	26	192	32	149	38	154	42	135
14	194	24	124	26	142	32	112	38	120	42	107	50	092
18	028	26	093	32	072	38	085	42	079	50	069	56	066
K=4		32	039	38	052	42	051	50	047	56	048	62	046
s	Q_w	38	0,024	42	0,029	50	0,027	56	0,030	62	0,031	71	0,026
8	0,431	42	0085	50	012	56	016	72	0099	78	010	86	012
14	273	50	0008	54	0081	62	0084	78	0048	86	0060	96	0075
18	125			56	0055	72	0036	86	0024	98	0029	104	0034
24	069			62	0017	78	0012	98	0009	104	0013	122	0013
26	042			72	0,0001	96	0,0003			114	0,0007	126	0,0008
32	0046												

N=4								N=5	
K=3		K=4		K=5		K=6		K=3	
s	Q_w	s	Q_w	s	Q_w	s	Q_w	s	Q_w
19	0,342	32	0,200	41	0,210	46	0,218	46	0,213
21	300	36	158	43	162	52	163	50	163
25	207	40	105	51	107	62	108	56	096
27	175	46	068	57	075	68	073	60	063
29	148	50	052	61	055	74	056	62	056
33	0,075	54	0,033	67	0,034	80	0,037	66	0,038
35	054	62	012	81	012	100	010	74	015
37	033	66	0062	85	0067	108	0061	78	0053
41	017	70	0027	93	0023	118	0028	82	0028
45	0017	74	0009	101	0014	128	0009	86	0009
				105	0,0006				

Пример 1.4 (продолжение):

$K=3$, $N=8$, $\widehat{w}(3)=0,497$.

$SL\approx 1-\Phi_7^{\chi^2}\left(3\cdot 7\cdot 0,497\right)=0,165\Rightarrow$ Совокупного согласия нет.

Замечание: можно показать, что среднее значение коэффициента корреляции Спирмена между всеми парами переменных

$$\bar{\bar{\rho}}_s=\frac{K\widehat{w}(K)-1}{K-1}. \qquad (7)$$

Если ранжировки согласованные, то итоговое ранжирование можно сделать на основании средних рангов объектов $\bar{r}^n$ (фактически по сумме мест).

Пример 1.6: при разработке квалиметрической методики экспертами ранжировались показатели из дерева свойств. Цель – приписать им коэффициенты весомости $v_n\geq 0,\ \sum_n v_n=1$.

Очевидно, их можно выбрать так: $v_n=\dfrac{2\bar{\bar{r}}-\bar{r}^n}{\sum_n\left(2\bar{\bar{r}}-\bar{r}^n\right)}$. (8)

1.3. Корреляционный анализ категоризованных переменных

Переменная (признак) называется категоризованной, если ее возможные значения описываются конечным числом состояний или градаций (категорий, уровней). Например, пол индивидуума (женский, мужской), социальное положение (рабочий, служащий, учащийся, безработный).

Иногда количественные признаки превращают в категоризованные. Например, возраст: детство (0 – 10 лет), юность (11 – 30 лет), зрелость (30 – 60 лет), старость (60 – 80 лет), долголетие (от 80 лет).

Категоризованные переменные могут встречаться вперемежку с количественными в таблице “объект-свойство”.

Пример 1.7: таблица объект-свойство

n	Рост, x_1	Пол, x_2	Вес, x_3	Социальное положение, x_4	Возраст, x_5
1	139	Ж	63	учащийся	Д
2	181	М	100	рабочий	З
3	⋮	М	⋮	рабочий	Ю
4		Ж		безработный	З
5		М		служащий	З
6		М		рабочий	Ю
7		Ж		безработный	С
8		М		безработный	До

Ясно, что работать с категоризованными переменными так же, как с количественными, нельзя. Например, какой смысл был бы у $\overline{x}_2$? Поэтому, при корреляционном анализе нескольких категоризованных переменных данные преобразуются к виду многовходовой таблицы сопряженности признаков.

Двухвходовая таблица сопряженности между x_2 и x_4

$x_{2i}\backslash x_{4j}$	У	Р	Б	С	$n_{i.} = \sum_j n_{ij}$
Ж	1	0	2	0	3
М	0	3	1	1	5
$n_{.j} = \sum_i n_{ij}$	1	3	3	1	8 = *N*

В ячейках таблицы сопряженности стоят числа n_{ij} объектов (из общего числа N обследованных), у которых значение первого признака на уровне i-й градации, второго – на уровне j-й градации.

Показатель степени тесноты статистической связи двух категоризованных переменных (вместо коэффициента корреляции Спирмена) выбирают так, чтобы эта характеристика принимала тем большее значение, чем больше анализируемая ситуация отличается от независимости исследуемых переменных.

Критерий независимости двух случайных величии

Пусть ξ и η – дискретные случайные величины,

x_1 , x_2 , …, x_r – возможные значения величины ξ,

y_1, y_2 , …, y_c – возможные значения величины η.

Проведено N наблюдений, ν_{ij} – случайные числа в ячейках таблицы сопряжённости.

Проверяется гипотеза:

H_0: ξ и η независимы (вероятность наблюдению попасть в ячейку ij $P\{(\xi = x_i) \cap (\eta = y_j)\} = P(\xi = x_i) P(\eta = y_j) = p_i p_j$) .

H_1: ξ и η зависимы, $P\{(\xi = x_i) \cap (\eta = y_j)\} \neq p_i p_j$, т. е. другое распределение.

Задача сводится к критерию согласия χ^2, каждая ячейка – разряд. Статистика критерия $H^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(\nu_{ij} - Np_i p_j\right)^2}{Np_i p_j}$.

$$SL = P\left\{H^2 \geq h^2 \mid H_0\right\} \approx 1 - \Phi^{\chi^2}_{(r-1)\cdot(c-1)}\left(h^2\right).$$

Поскольку параметры p_i и p_j неизвестны, то они заменяются их оценками: $\widehat{P}_i = \frac{\sum_j \nu_{ij}}{N}$ и $\widehat{P}_j = \frac{\sum_i \nu_{ij}}{N}$. Так как $\sum_i \widehat{P}_i = 1$ и $\sum_j \widehat{P}_j = 1$, то число независимых параметров равно (r – 1) + (c – 1) и число степеней свободы распределения χ^2

$$r \cdot c - 1 - \left(r + c - 2\right) = \left(r-1\right)\cdot\left(c-1\right).$$

Определение: коэффициентом сопряжённости Крамера называется величина

$$V = \left[\frac{h^2}{N \min\left(r-1, c-1\right)}\right]^{\frac{1}{2}},$$

где h^2 – значение статистики из критерия χ^2,

r – количество категорий одной переменной,

c – количество категорий другой переменной.

Свойства:

1. $0 \leq V \leq 1$,
2. $V = 0 \Rightarrow$ строгая статистическая независимость,
3. $V = 1 \Rightarrow$ возможность восстановления одной величины по другой.

Пример 1.7 (продолжение):

$$h^2 = N\left[\sum_{i=1}^{r}\frac{1}{n_{i.}}\sum_{j=1}^{c}\frac{n^2{}_{ij}}{n_{.j}} - 1\right], \qquad (9)$$

$$h^2 = 8\left[\frac{1}{n_{1.}}\left(\frac{n_{11}^2}{n_{.1}}+\frac{n_{12}^2}{n_{.2}}+\frac{n_{13}^2}{n_{.3}}+\frac{n_{14}^2}{n_{.4}}\right)+\frac{1}{n_{2.}}\left(\frac{n_{21}^2}{n_{.1}}+\frac{n_{22}^2}{n_{.2}}+\frac{n_{23}^2}{n_{.3}}+\frac{n_{24}^2}{n_{.4}}\right)-1\right]=$$

$$=8\left[\frac{1}{3}\left(\frac{1}{1}+\frac{0}{3}+\frac{4}{3}+\frac{0}{1}\right)+\frac{1}{5}\left(\frac{0}{1}+\frac{9}{3}+\frac{1}{3}+\frac{1}{1}\right)-1\right]=5{,}16,$$

$$SL = 1-\Phi_{1\cdot 3}^{\chi^2}(5{,}16) = 0{,}161.$$

Принимаем гипотезу H_0, т. е. пол и социальное положение независимы.

$V = \left[\frac{5{,}16}{8\cdot 1}\right]^{\frac{1}{2}} = 0{,}803$ – довольно большое значение, которое, однако, не гарантирует зависимости!

Если таблица трех- или более входовая, то уровень значимости рассчитывается аналогично. Например, пол, социальное положение, возраст. $r = 2$, $c = 4$, $s = 5$.

$SL = 1-\Phi_{31}^{\chi^2}(41{,}3) = 0{,}102$ – уровень значимости данных против гипотезы независимости всех трех величин. Такое значение интерпретируется: "данные согласуются с этой гипотезой".

Переменная множественного отклика

Пример 1.8 (развитие 1.7): опрашиваемых попросили указать любимые напитки (до трех) из списка:

1. Пепси-кола
2. Пиво
3. Сок
4. Кока-кола
5. Швепс
6. Кофе.

При обработке в Statistica ответы помещают в соседние 3 столбца:

	1 VAR1	2 VAR2	3 VAR7	4 VAR8	5 VAR9
1	f	y	пепси	кока	кофе
2	m	p	пиво	швепс	
3	m	p	сок	пиво	швепс
4	f	b	кока	пепси	
5	m	c	швепс	пиво	кофе
6	m	p	пиво	пепси	швепс
7	f	b	сок	кофе	
8	m	b	кофе		

Var7, Var8, Var9 – переменная множественного отклика (набор (set) множественного отклика).

Важно, чтобы одинаковые метки (слова) имели одинаковые числовые коды. Для этого лучше начинать с кодов, а затем связывать с метками (через Apply To...).

Изучим связь между полом и любимыми напитками. Statistics→Basic Statistics and tables→ Multiple Response Tables Specify tables (select variables). Set 1 – Var1, Set 2 – (Var7 – Var9) →OK→Codes All →OK.

Ignore multiple identical responses означает, что, например, ответ пиво/пиво/ воспринимается как пиво/ /.

Можно посмотреть характеристики самой переменной: Frequency tables.

	Frequencies (Identical resp. were ignored) Variable: VAR7 (Multiple Response Variable)			
N=8 Category	**Count**	Prcnt.of Responses	Prcnt.of Cases	
пепси	3	15,79	37,50	
пиво	4	21,05	50,00	
сок	2	10,53	25,00	
кока	2	10,53	25,00	
швепс	4	21,05	50,00	
кофе	4	21,05	50,00	
Totals	19	100,00	237,50	

19 – столько всего было ответов.

4/19 =0,2105. Значит, из всех упоминаний 21,05% – пиво.

4/8 = 0,5. Значит, у 50% опрошенных упоминалось пиво.

Summary: Review summary tables – таблица сопряженности.

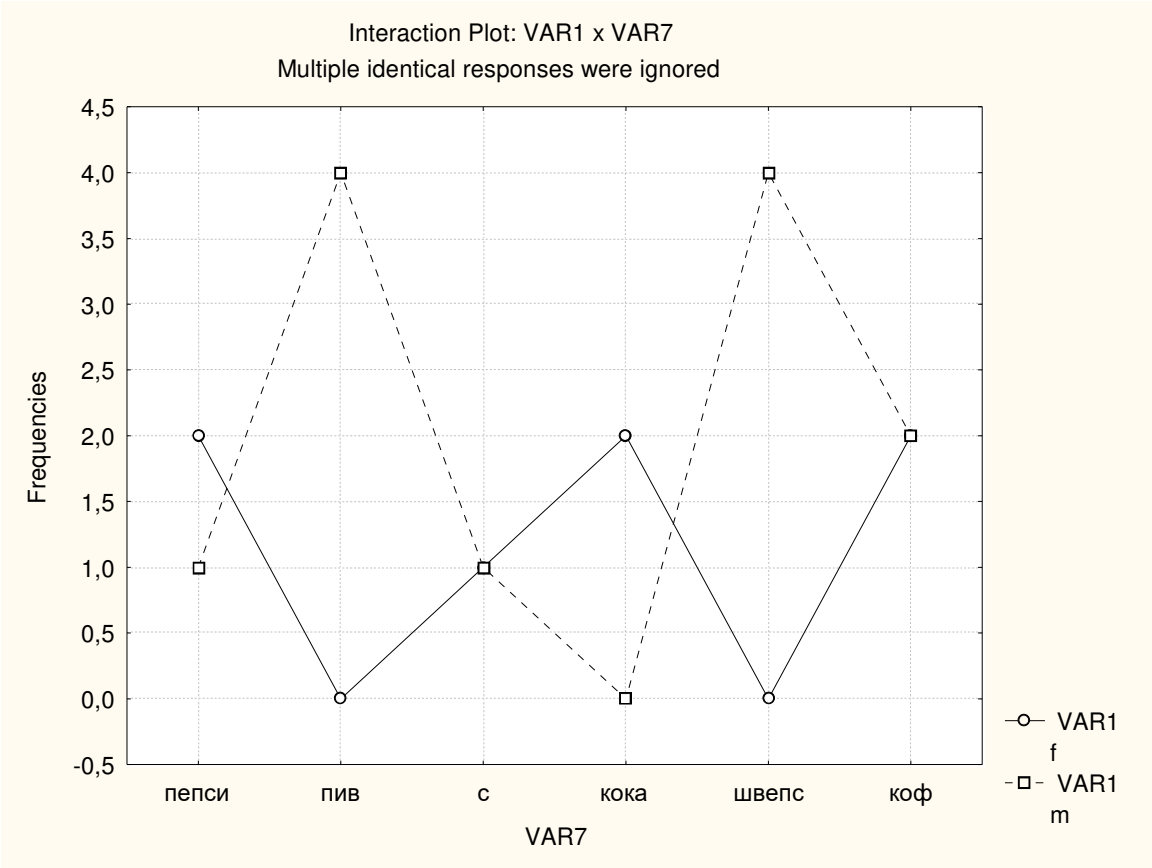
Summary Table for all Multiple Response Items (полСоцполНа
Totals/percentages based on number of respondents
Multiple identical responses were ignored

N=8 VAR1	VAR7 пепси	VAR7 пиво	VAR7 сок	VAR7 кока	VAR7 швепс	VAR7 кофе	Row Totals
f	2	0	1	2	0	2	3
m	1	4	1	0	4	2	5
All Grps	3	4	2	2	4	4	8

Выбирая в Options Percentage of..., Totals/percentages based on number of respondents, Totals/percentages based on number of responses, можно в таблице сопряжённости получить ещё и разнообразные проценты.

Например, какой процент женщин пьёт кофе? 2/3 даёт 66.67%. Сколько кофе выпивается женщинами? 50%. Сколько процентов из выпитого женщинами составляет кофе? 28.57%.

Interaction plots of frequencies – наглядное изображение таблицы сопряжённости.



1.4. Упражнения и лабораторный практикум

Упражнение 1.1.

Доказать теорему (2), используя формулу

$$D(X_i + X_j) = DX_i + DX_j + 2\operatorname{cov}(X_i, X_j)$$

и формулу несмещенной и состоятельной оценки дисперсии.

Упражнение 1.2.

Проверить формулы для оценки ковариационной матрицы (векторную и через матрицу объект-свойство (3)).

Упражнение 1.3.

По данным примера 1.1 в MathCAD найти оценки ковариационной и корреляционной матриц, множественный коэффициент корреляции. Применить критерий Уилкса – Бартлетта.

Упражнение 1.4.

По данным примера 1.2 проверить гипотезу о коэффициенте корреляции. Каков бы был *SL* при 10 наблюдениях при той же оценке коэффициента корреляции?

Упражнение 1.4а.

Чему равен множественный коэффициент корреляции в случае всего двух переменных?

Упражнение 1.4б.

Вывести формулу (4.2).

Упражнение 1.4в.

Вывести формулу (4.3).

Упражнение 1.5.

Повторить в Statistica примеры 1.1 и 1.2.

Корреляционная матрица: Statistics -> Basic Statistics and Tables-> Correlation Matrices -> One variable list…Options: Display r, p-levels, and N's.

Частные коэффициенты корреляции: Advanced, Partial correlations, в левый столбец – пара переменных, в правый – исключаемые влияния.

Множественный коэффициент корреляции и тест Уилкса – Бартлетта: Statistics -> Multivariate Exploratory Techniques -> Principal Components & Classification Analysis, Variables, Ok, Descriptives Tab, Correlation matrix Inverse. Здесь же логарифм определителя оценки корреляционной матрицы. Подтвердить результат для множественного коэффициента корреляции через коэффициент детерминации: Statistics -> Multiple Regression. Найти SL теста Уилкса – Бартлетта в примере 1.1.

Упражнение 1.6.

Подобрать данные для корреляционного анализа 3 – 4 количественных переменных (10 – 15 наблюдений), содержащие ложную корреляцию. Например, располагаемый личный доход, цена и спрос на некоторый продукт (Доугерти К. Введение в эконометрику. М.: Инфра – М, 2001. с. 52). Оценить корреляционную матрицу, полные и частные коэффициенты корреляции, значимость их отличий от нуля, множественный коэффициент корреляции. Выявить ложные корреляции или маскировку ложной корреляцией должной. Для этого обратить особое внимание на возможное изменение знака и значимости при переходе к частным коэффициентам корреляции. Объяснить механизм ложной корреляции, проиллюстрировать возникновение ложной корреляции диаграммой, подобной примеру 1.3, сделать выводы. Проверить критерием Уилкса – Бартлетта гипотезу о попарной независимости всех компонент случайного вектора.

Упражнение 1.6а.

Быстрая походка, особенно в пожилом возрасте, спасает от фатальных сердечных заболеваний. Люди, которые медленно ходят, почти втрое чаще могут умереть от заболеваний сердца, чем те, кто передвигается споро, живо. К такому выводу пришла группа французских ученых, результаты исследования которыми 3 тыс мужчин и женщин с помощью видеокамер, фиксирующих скорость, поместил Британский медицинский журнал.

Обсудить на предмет ложной корреляции. Как следовало производить исследование?

Упражнение 1.7.

Показать формулу $\widehat{w}(K) = \dfrac{12}{(N^3 - N)} \sum_{n=1}^{N} (\overline{r}^n - \overline{\overline{r}})^2$.

Рассмотреть случай полного согласия, полного отсутствия согласия.

Получить результаты примера 1.4 (продолжение).

Упражнение 1.8.

6 экспертов ранжировали 4 объекта. *w*^ = 0.444. Согласованны ли мнения экспертов? Найти точный и приближенный уровни значимости.

Упражнение 1.9.

Повторить в Statistica пример 1.4:

1) получить матрицу коэффициентов корреляции Спирмена с уровнями значимости. Nonparametrics -> Correlations (Spearman, Kendall tau, gamma);
2) получить коэффициент конкордации с уровнем значимости. Nonparametrics -> Comparing multiple dep. samples (variables). Данные должны быть предварительно транспонированы (Data -> Transpose);
3) проверить формулу (7).

Упражнение 1.10.

Проверить свойства весов (8): min значение, сумму. Найти веса, используя данные примера 1.4.

Упражнение 1.12.

Вывести формулу (9) для статистики h^2 при оценке гипотетических вероятностей по данным.

Упражнение 1.13.

Утверждается, что результат действия лекарства зависит от способа его применения (*A, B, C*). Так ли это?

Исход \ Способ	*A*	*B*	*C*
Благоприятный	11	17	16
Неблагоприятный	20	23	19

Упражнение 1.14.

Повторить в Statistica пример 1.7:

1. Statistics→Basic Statistics and tables→Tables and Banner→Analysis: Cross tabulation tables→Specify tables.
2. Выбрать, например, x_2 в List 1 и x_4 в List 2.
3. Выбрать Use all integer codes… .
4. Review summary tables→Ok→Summary Frequency tables.
5. В Options выбрать Expected frequencies и нужные статистики.
6. На экране появится таблица сопряженности.
7. Нажать Detailed two-way tables.
8. Появится статистика χ^2, *SL* и коэффициент сопряженности Крамера.
9. Добавить переменную «возраст».

a. Найти SL и коэффициент сопряжённости Крамера для «социальное положение – возраст».

b. Провести анализ, включая исследование зависимостей различных пар переменных в различных категориях третьей переменной, описать суть выявленных зависимостей. Проверить значимость выводов (данных). В частности, рассмотреть ситуацию в категории людей зрелого возраста.
c. Проверка гипотезы о независимости всех факторов. Найти SL. Почему d.f. = 31?

Упражнение 1.15.

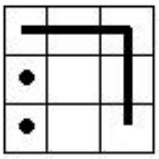
Повторить в Statistica пример 1.8. Получить таблицы частот и сопряженности, график взаимодействий. Выбирая в Options Percentage of..., Totals/percentages based on number of respondents, Totals/percentages based on number of responses, получить еще и разнообразные проценты. Истолковать проценты, связанные с соком: 33.33%, 20%, 25%, 50%, 14.29%, 8.33%,10.53%.

Упражнение 1.16.

Подобрать данные 2-х категоризованных переменных, одна из которых множественного отклика. Получить таблицу частот, сопряженности, график взаимодействий. Выбирая в Options Percentage of..., Totals/percentages based on number of respondents, Totals/percentages based on number of responses, получить и истолковать еще и разнообразные проценты (подобно процентам, связанным с соком в лекционном примере). По возможности проверить значимость выводов (данных).

2. Распознавание образов

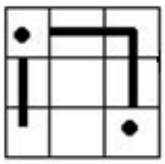
Пример 2.1: вы написали букву П и со сканировали ее. В компьютере проявилась картинка



Вы запускаете программу распознавания текста. Она видит

$\vec{x} = (1, 1, 1, 0.5, 0, 1, 0.5, 0, 1)^T$ – информативные признаки (образ).

Могло со сканироваться



$\vec{x}'$ – другой образ, но тоже буквы П. То есть это образы из класса образов буквы П.

Задача распознавания: по образу $\vec{x}$ выработать оценку $\hat{y}$ номера класса образов, к которому $\vec{x}$ принадлежит.

Пример 2.1 (продолжение): желательно выработать $\hat{y} = 16$ – номер прообраза (буквы П) в алфавите. Программа могла ошибиться (решить, что это буква О) и выработать $\hat{y} = 15$.

2.1. Распознавание образов с известными априорными вероятностями

Пусть $p_1,\ p_2, \ldots, p_q, \ldots,\ p_Q$ – априорные вероятности предъявления (появления) классов $1, 2, \ldots, q, \ldots, Q,\ p_q = P\{Y = q\}$ известны.

Известны также $p_{\vec{X}}\left(\vec{x} \middle| Y = q\right) = p\left(\vec{x} \middle| q\right)$ – условные плотности распределения наблюдаемой величины $\vec{X}$ при условии, что был предъявлен класс $Y = q$.

Именно различие этих функций делает возможным распознавание.

Так как класс предъявляется случайный, наблюдаемая величина $\vec{X}$ случайна, то и оценка, вырабатываемая по ней, тоже случайна, $\hat{Y}$.

Правильному решению очевидно соответствует $\hat{Y} = Y$.

Желательно $P\left\{\hat{Y} = Y\right\} = p \to \max$. (10)

Определение: оптимальная модель распознавания (оптимальное решающее правило):

1. Пространство значений Ω величины $\vec{X}$ разбивается на Q непересекающихся частей A_1, A_2, …, A_Q.

2. Объявляется класс $\hat{Y} = q$ при попадании $\vec{X}$ в область A_q.

3. Границы между областями A_q выбираются так, чтобы вероятность правильного решения была максимальной (10).

Найдем области A_q.

Обозначим $\alpha_{kq} = P\left\{\hat{Y} = k \middle| Y = q\right\} = P\left\{\vec{X} \in A_k \middle| q\right\} = \int\limits_{A_k} p\left(\vec{x}\middle| q\right) d\vec{x}$.

Полная вероятность правильного распознавания:

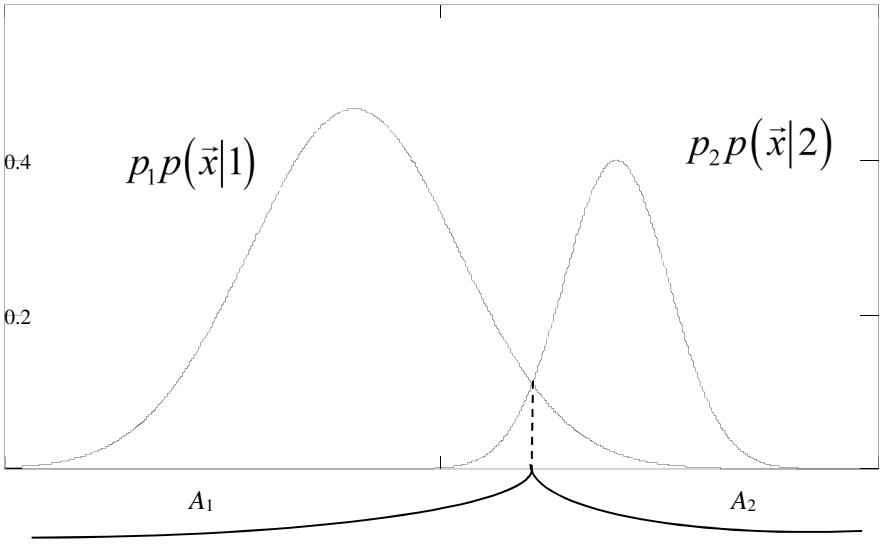
$$p = P\left\{\hat{Y} = Y\right\} = \sum_{q=1}^{Q} P\{Y = q\} P\left\{\hat{Y} = Y \middle| Y = q\right\} = \sum_{q=1}^{Q} p_q \alpha_{qq},$$

или $p = \sum\limits_q p_q \int\limits_{A_q} p\left(\vec{x}\middle| q\right) d\vec{x} = \sum\limits_q \int\limits_{A_q} p_q p\left(\vec{x}\middle| q\right) d\vec{x} \to \max$.

Очевидно, под интегрирование каждого выражения $p_q p\left(\vec{x}\middle| q\right)$ нужно отвести те места или точки области Ω, где эта подынтегральная функция больше всех других (при других q'):

$$A_q = \left\{\vec{x} : p_q p\left(\vec{x}\middle| q\right) = \max_{1 \le l \le Q} p_l p\left(\vec{x}\middle| l\right)\right\}.$$

Пример 2.2: графическое изображение областей. X - скалярная, $Q = 2$.



Очевидно, построенное решающее правило можно переформулировать: если наблюдалось $\vec{X} = \vec{x}^0$, то

$$\hat{y} = \arg\max_{1 \le q \le Q} p_q p\left(\vec{x}^0 \middle| q\right). \qquad (11)$$

Распознавание в модели Фишера

Определение (модель Фишера): распределение вектора $\vec{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_r \end{pmatrix}$ информативных признаков для класса q является нормальным с математическим ожиданием $M\vec{X} = \vec{a}^q$ и ковариационной матрицей $K = K_{\vec{X}} = M\left(\vec{X} - \vec{a}^q\right)\left(\vec{X} - \vec{a}^q\right)^T$, одинаковой для всех классов. Следовательно,

класс от класса отличается центром.

$$p\left(\vec{x}|q\right) = \frac{1}{(2\pi)^{\frac{r}{2}}\sqrt{\det(K)}} \exp\left\{-\tfrac{1}{2}\left(\vec{x} - \vec{a}^q\right)^T K^{-1}\left(\vec{x} - \vec{a}^q\right)\right\}.$$

Решающее правило (11) будет иметь вид

$$\hat{y} = \arg\max_q \ln p_q p(\vec{x} \mid q) = \arg\max_q \left\{ \ln p_q + \ln \left[(2\pi)^{\frac{r}{2}} \sqrt{\det(K)} \right]^{-1} - \right.$$

$$\left. - \frac{\left(\vec{x} - \vec{a}^q\right)^T K^{-1} \left(\vec{x} - \vec{a}^q\right)}{2} \right\} = \arg\max_q \left\{ \ln p_q - \frac{1}{2} \vec{x}^T K^{-1} \vec{x} + \frac{1}{2} \vec{x}^T K^{-1} \vec{a}^q + \right.$$

$$\left. + \frac{1}{2} \left(\vec{a}^q\right)^T K^{-1} \vec{x} - \frac{1}{2} \left(\vec{a}^q\right)^T K^{-1} \vec{a}^q \right\} = \arg\max_q \left\{ \left(\vec{a}^q\right)^T K^{-1} \vec{x} + \ln p_q - \right.$$

$$\left. - \frac{1}{2} \left(\vec{a}^q\right)^T K^{-1} \vec{a}^q \right\}.$$ Окончательно:

$$\hat{y} = \arg\max_q g_q(\vec{x}), \tag{12}$$

где $g_q(\vec{x}) = \left(\vec{A}^q\right)^T \vec{x} + b_q$ – линейная классифицирующая функция, в которой

$$\vec{A}^q = K^{-1} \vec{a}^q, \qquad b_q = -\frac{1}{2} \left(\vec{A}^q\right)^T \vec{a}^q + \ln p_q .$$

2.2. Дискриминантный анализ (классификация с обучающими выборками)

По имеющемуся образу (или наблюдению) требуется определить номер класса реализаций, к которому он принадлежит. Считаем справедливой модель Фишера, но теперь $p_q, \vec{a}^q, K$ не знаем.

Мы сможем оценить эти параметры, если имеются наблюдений, про которые известно, к какому классу они принадлежат (обучающие выборки):

$$\vec{x}^{n_q} = \begin{pmatrix} x_1^{n_q} \\ ..... \\ x_r^{n_q} \end{pmatrix} \quad n_q \in 1,...,N_q;\ q \in 1,...,Q .$$

Тогда

$\hat{\vec{a}}^q = \bar{\vec{x}}^q = \frac{1}{N^q}\sum_{n_q=1}^{N_q} \vec{x}^{n_q}$. Оценку ковариационной матрицы можно получить по обучающей выборке для любого класса

$\hat{K}^q = \frac{1}{N_q - 1}\sum_{n_q=1}^{N_q} (\vec{x}^{n_q} - \bar{\vec{x}}^q)(\vec{x}^{n_q} - \bar{\vec{x}}^q)^T$, или сразу по всем обучающим данным

$\hat{K} = \frac{\sum_{q=1}^{Q} (N_q - 1)\hat{K}^q}{\sum_{q=1}^{Q} N_q - Q}$. Оценка априорных вероятностей $\hat{p}_q = \frac{N_q}{\sum_{q=1}^{Q} N_q}$.

Заменяя $p_q, \vec{a}^q, K$ в решающем правиле (12) этими оценками, получаем «подстановочное» решающее правило. Можно классифицировать новое наблюдение, не входившее в обучающие выборки.

Пример 2.3: имеются данные по 7 предприятиям машиностроительного комплекса:

x_1 – фондоотдача основных производственных фондов, руб,

x_2 – затраты на рубль произведённой продукции, коп,

x_3 – затраты сырья и материалов на один рубль продукции, коп,

y – код группы успешности.

n	x_1	x_2	x_3	y
1	0,5	94	8,5	1
2	0,67	75,4	8,79	1
3	1,2	93,8	6,95	2
4	0,68	85,2	9,1	1
5	1,52	81,5	4,95	2
6	1,46	86,5	4,7	2
7	0,55	98,8	8,47	1

Необходимо классифицировать новое (8-е) предприятие, имеющее следующие показатели:

$$\vec{x}^8 = \begin{pmatrix} 0,99 \\ 84 \\ 4,85 \end{pmatrix}.$$

Находим оценки $\hat{\vec{a}}^1 = \frac{1}{4}\left[\begin{pmatrix} 0,5 \\ 94 \\ 8,5 \end{pmatrix} + ... + \begin{pmatrix} 0,55 \\ 98,8 \\ 8,47 \end{pmatrix}\right] = \begin{pmatrix} 0,6 \\ 88,4 \\ 8,72 \end{pmatrix}$,

$\hat{\vec{a}}^2$ – аналогично.

$$\hat{K} = \begin{pmatrix} 0,016 & -0,86 & -0,067 \\ -0,86 & 79,06 & 1,46 \\ -0,067 & 1,46 & 0,66 \end{pmatrix},$$

$$\vec{A}^1 = \hat{K}^{-1}\hat{\vec{a}}^1 = \begin{pmatrix} 725,1 \\ 7,7 \\ 69,4 \end{pmatrix},$$

$$\hat{p}^1 = \frac{4}{7} = 0,571, \quad \hat{p}^2 = \frac{3}{7} = 0,4285, \quad b_1 = -860,98.$$

$$g_1(\vec{x}) = 725{,}1x_1 + 7{,}7x_2 + 69{,}4x_3 - 860{,}98,$$

$g_2(\vec{x})$ – аналогично. $g_1(\vec{x}^8) = 840{,}7$.

Вычисляя $g_2(\vec{x}^8)$, находим: $g_1(\vec{x}^8) > g_2(\vec{x}^8)$, значит, относим 8-е наблюдение к 1-му классу.

Можно найти SL = 0,0069 – значимость данных против гипотезы о неразличимости классов.

2.3. Автоматическая классификация (кластер-анализ)

Допустим, что у нас нет ни априорных вероятностей $p_1,\ p_2, \ldots, p_q, \ldots,\ p_Q$ появления образов классов, ни информации о распределении вектора наблюдений внутри классов $p\left(\vec{x}\middle|q\right)$. Нет и обучающих выборок. Исходная информация о классифицируемых объектах представлена в виде матрицы "объект – свойство":

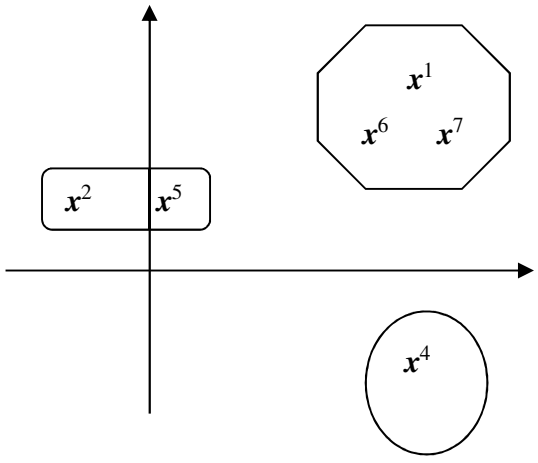
$$Z = \begin{pmatrix} \vec{x}^{1^T} \\ ..... \\ \vec{x}^{N^T} \end{pmatrix} = \begin{pmatrix} x_1^{\,1} & \ldots & x_r^{\,1} \\ ..... & & \\ x_1^{\,N} & \ldots & x_r^{\,N} \end{pmatrix},$$

x_j^n – значение j-го признака на n-м обследованном объекте.

Задача классификации состоит в том, чтобы всю анализируемую совокупность N объектов разбить на сравнительно небольшое число Q (заранее известное или нет) однородных, в определенном смысле, групп или классов.

Полученные в результате разбиения классы называют также кластерами, таксонами.

Пример 2.4: разбиение объектов на кластеры.



$Q = 3,\ N = 7,\ r = 2.$

$S^1 = \left\{\vec{x}^1, \vec{x}^6, \vec{x}^7\right\},\ S^2, S^3$ – кластеры.

Ясно, что классификация производится на основании расстояния $d(\vec{x}^m, \vec{x}^n)$, близкие относим к одному классу, далекие – к разным. В качестве расстояния выбирают обычное евклидово расстояние

$$d_E(\vec{x}^m, \vec{x}^n) = \sqrt{(x_1{}^m - x_1{}^n)^2 + \left(x_2{}^m - x_2{}^n\right)^2 + \ldots + (x_r{}^m - x_r{}^n)^2}$$

или взвешенное евклидово

$$d_{WE}(\vec{x}^m, \vec{x}^n) = \sqrt{w_1(x_1{}^m - x_1{}^n)^2 + w_2\left(x_2{}^m - x_2{}^n\right)^2 + \ldots + w_r(x_r{}^m - x_r{}^n)^2},$$

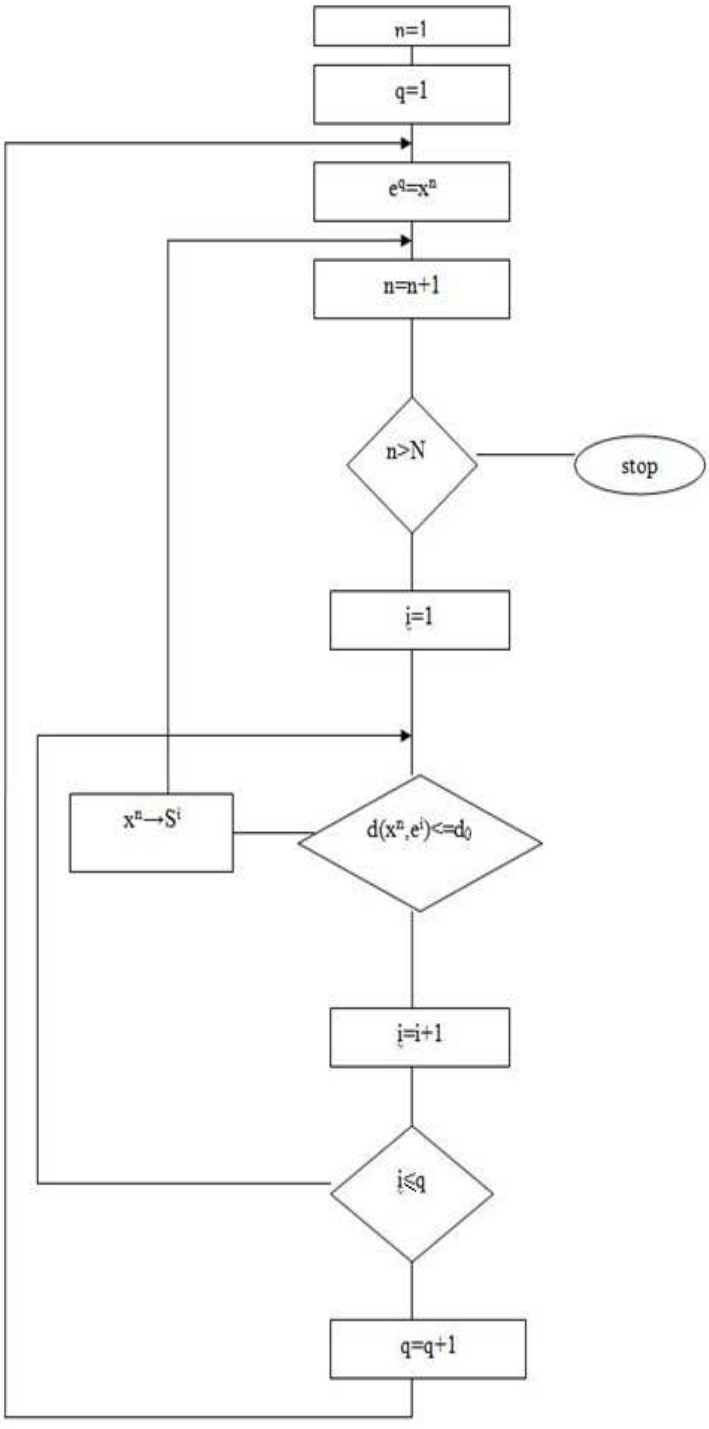
$$0 \le w_i \le 1, \qquad \sum w_i = 1.$$

Веса выбираются пропорционально степени важности компоненты x_i для отнесения объекта к тому или иному классу.

Процедуры, в которых элементы выборки поступают на классификацию по одному – последовательные, а в которых выборка вся сразу – параллельные.

Некоторые конкретные алгоритмы

1. Простой последовательный (должен быть задан радиус классов d_0). Блок-схема алгоритма:



Берется первый элемент,
q – номер класса,
объект x^n объявляется ядром e^q q-го класса,

«примеряем» новый объект n к классу i,

пробуем другой открытый класс,

открываем новый класс,
n-й объект не подошел ни к одному уже открытому классу.

2. Алгоритм K-средних (параллельный), число классов Q должно быть задано.

- $m = 0$ (номер разбиения). Строим каким-либо образом начальное разбиение $S_0 = (S_0^{\,1}, ..., S_0^{\,Q})$.

- Пусть уже построено m-е разбиение $S_m = (S_m{}^1, ..., S_m{}^Q)$. Вычислим в каждом из этих классов средние: $\vec{e}_m{}^q = \frac{1}{N_{S_m^q}} \sum_{\vec{x}^n \in S_m^q} \vec{x}^n$.

- Строим минимально-дистанционное разбиение, порождаемое набором $\left\{\vec{e}_m{}^q\right\}$, по формулам:

$$S_{m+1}^q = \left\{ \vec{x}^n : q = \arg\min_{1 \le q' \le Q} d(\vec{x}^n, \vec{e}_m{}^{q'}) \right\},$$

то есть каждый элемент относим к тому классу, к центру которого он ближе.

- $S_{m+1} \neq S_m \Rightarrow m = m+1, \;\; S_{m+1} = S_m \Rightarrow S_m = S_*, \text{stop}$.

Замечания:

1. Можно показать, что для любого начального разбиения S_0 алгоритм через конечное число шагов заканчивает работу.

2. В ряде случаев начальное разбиение S_0 задается как минимально-дистанционное, порожденное некоторым набором точек $(\vec{e}_{-1}{}^1, ..., \vec{e}_{-1}{}^Q)$, в качестве которых, например, можно взять Q наиболее отстоящих друг от друга наблюдений $(\vec{x}^{n_1}, ..., \vec{x}^{n_Q})$ или первые Q наблюдений. Результат классификаций зависит от выбора S_0, поэтому для проверки устойчивости результата рекомендуется варьировать выбор.

3. Если из априорных сведений нельзя сразу выбрать число классов Q, его находят путем перебора.

2.4. Упражнения и лабораторный практикум

Упражнение 2.1.

Ваш друг ежедневно заходит в игровой салон (с вероятностью 0,85), где играет в компьютерные игры, играя "до отказа", или заходит в клуб (с вероятностью 0,15), где беседует с друзьями. В салоне в среднем проводит 1 час, в клубе 3 часа со стандартным отклонением 1 час. Сегодня он затратил 2,5 часа. Где он был? Построить области A_1 и A_2.

Упражнение 2.2.

Показать, что решающее правило максимальной апостериорной вероятности (называть класс, гипотеза о котором имеет максимальную апостериорную вероятность при наблюдаемом $\vec{x}^0$) эквивалентно правилу (11). Найти апостериорные вероятности в упражнении 2.1.

Упражнение 2.3.

Повторить в Statistica пример 2.3.

Multivariate Exploratory Techniques ->Discriminant Analysis -> Variables... Grouping... Independent... Codes(1-2)... ->.

Найти оценки средних и ковариационной матрицы: Advanced options (Stepwise analysis). Descriptive Tab.

Получить выражение для линейной классифицирующей функции 2-й группы, классифицировать 8-е наблюдение, вычислив значения линейных классифицирующих функций.

Классифицировать все наблюдения и сравнить классификацию обучающих наблюдений с фактической. Проверить гипотезу о неразличимости классов. Classification Tab -> Classification functions... -> Classification of cases (будут расклассифицированы все наблюдения в файле данных, включая обучающие, в порядке предпочтительных классификаций).

Упражнение 2.4.

Для распознавания предъявляется изделие с контролируемым размером a_1 или a_2 ($a_1 > a_2$). Априорные вероятности предъявления – p_1 и p_2 соответственно. Для распознавания проводится N независимых измерений изделия прибором с погрешностью σ. Указать классы образов. Построить оптимальное решающее правило распознавания (указать A_1 и A_2). Найти вероятности ошибок α_{12} и α_{21} и полную вероятность ошибки распознавания. Можно ли применить построенное правило к проверке партий готовых изделий с номиналами a_1 – годная, и a_2 – брак? Можно ли N-кратным измерением одного изделия принять решение о том, взято оно из партии с номиналом a_1 или a_2?

Упражнение 2.5.

Расклассифицировать 6 последовательных наблюдений, $d_0 = 2$.

$\{\vec{x}^n\}: 4,3,1,6,2,0$. А в обратном порядке?

Упражнение 2.6.

Запрограммировать в MathCAD простой последовательный алгоритм и повторить упражнение 2.5.

Упражнение 2.7.

Алгоритмом K-средних расклассифицировать наблюдения из предыдущего упражнения в 2 класса.

Упражнение 2.8.

Повторить в Statistica предыдущее упражнение, методами дисперсионного анализа проверить гипотезу о существенности различий между центрами классов (каков SL?).

Statistics → Multivariate Exploratory Techniques→ Cluster Analysis→ K-Means Clustering →Initial cluster centers →Choose the first N (Number of clusters) observations →Members of each cluster & distances →Analysis of variance.

Упражнение 2.9.

Методом *K*-средних произвести классификацию 7 предприятий машиностроительного комплекса из примера 2.3 по трём переменным – X_1 , X_2 , X_3 в 2 класса.

Проверить значимость данных по каждой переменной для классификации. Какая самая значимая, каков SL? Сравнить это с *Graph of means*. Стандартизировать данные. Сравнить итоговую классификацию с кодами «успешности», сделать экономические выводы. Что изменилось? Почему? Попробовать разбить данные на 3 – 4 класса, проверить значимость. Построить 3D – график наблюдений, повернуть его наглядно, обвести полученные кластеры. Истолковать результаты.

Graphs → 3D XYZ Graphs → Scatterplots. → VAR 1, 2, 3 → Options1 → Display Options → Case labels → Case names → OK → Polygon (рисовать).

3. Отбор наиболее информативных показателей и снижение размерности пространства признаков

В практической статистической работе число признаков *r*, регистрируемых на каждом из *N* объектов (стран, городов, семей и т. п.), очень велико (порядка 100). Имеющиеся многомерные наблюдения $\vec{x}^n \quad (n = 1, ..., N)$, занимающие большой объем, желательно заменить *N* наборами вспомогательных переменных $\vec{f}^n = \begin{pmatrix} f_1^n \\ ..... \\ f_{r'}^n \end{pmatrix}$ с числом $r' << r$ без существенной потери информации.

Предпосылки к этому: дублирование информации, доставляемой сильно взаимосвязанными признаками, не информативность признаков, мало меняющихся от объекта к объекту.

Цели:

1) визуализация исходных данных (если $r' = 1, 2$ или 3),

2) классификация объектов пойдет легче, если признаков будет меньше (r', а не r),

3) использование f_i в качестве предикторов (регрессоров) вместо x_i дает свои преимущества.

3.1. Метод главных компонент

Здесь пока нет цели уменьшить число новых переменных, т.е. $r' = r$. Просто лучше бы f_i были не коррелированы (т. е. каждая из них была бы самоценной). Этого можно достичь.

Будем считать переменные центрованными (т. е. $MX_i = 0$). Переход от исходных переменных $\tilde{X}_i$ к центрованным осуществляется так: $X_i = \quad \tilde{X}_i - \bar{\tilde{X}}_i$.

Ковариационная матрица – $K_{\vec{X}} = M\vec{X}\vec{X}^T$. $K_{\vec{X}} \neq diag$, то есть переменные X_i – зависимые случайные величины.

Найдём новые переменные X_i' – линейные комбинации старых: $\vec{X}' = P^{-1}\vec{X}$, которые будут некоррелированными $K_{\vec{X}'} = diag$. Какова матрица P^{-1}?

Теорема (о сингулярном разложении симметричной матрицы):

$$K_{\vec{X}} = S\Lambda S^T, \qquad (13)$$

$\Lambda = \begin{pmatrix} \lambda_1 & ... & 0 \\ 0 & ... & \lambda_r \end{pmatrix}$ – матрица из собственных значений,

$S = (\vec{e}_0^{\,1}, ..., \vec{e}_0^{\,r})$ – матрица из ортонормированных собственных векторов матрицы $K_{\vec{X}}$, т. е. $K_{\vec{X}} \cdot \vec{e}_0^{\,i} = \lambda_i \cdot \vec{e}_0^{\,i}$, λ_i – собственное значение, соответствующее собственному вектору $\vec{e}_0^{\,i}$.

$$S^{-1} = S^T. \qquad (14)$$

Следствие:

$\Lambda = S^T K_{\vec{X}} S = S^T M\vec{X}\vec{X}^T S = M\vec{X}'\vec{X}'^T$, где $\vec{X}' = S^T\vec{X}$, то есть

$$K_{\vec{X}'} = \Lambda \qquad (15)$$

и значит новые компоненты не коррелированы. Т. о. $P^{-1} = S^T$.

Если потребовать "сверхновых" компонент $\vec{X}''$: $K_{\vec{X}''} = I$, то переход к ним

$$\vec{X}'' = \Lambda^{-\frac{1}{2}} \cdot \vec{X}' \qquad (16)$$

Определение: координаты $\vec{X}'$ – главные компоненты, $\vec{X}'' = \vec{F}$ – нормированные главные компоненты (главные факторы). $A = S \cdot \Lambda^{\frac{1}{2}}$ называется матрицей нагрузок главных факторов на исходные признаки.

$$\vec{X} = A \cdot \vec{F}.$$

Свойства главных компонент

1. $M\vec{F} = MA^{-1}\vec{X} = A^{-1}M\vec{X} = A^{-1}\vec{0} = \vec{0}$, т. е. главные факторы центрованные, и так как $DF_i = 1$, нормированные.

2. $\sum_{i=1}^{r} DX_i = SpK_{\vec{X}} = Sp(S\Lambda S^T) = Sp(S^T S\Lambda) = Sp\Lambda = \sum_{i=1}^{r}\lambda_i$,

т. е. суммарная дисперсия совокупности признаков равна сумме собственных чисел, и о первых r' главных компонентах с наибольшими собственными числами (дисперсиями) можно сказать, что они учитывают долю $\frac{\sum_{i=1}^{r'}\lambda_i}{\sum_{i=1}^{r}\lambda_i}$ полной дисперсии.

3. Теорема Терстоуна.

$$K_{\vec{X}} = AA^T. \qquad (17)$$

4. $A^T A = \Lambda^{\frac{1}{2}} S^T S \Lambda^{\frac{1}{2}} = \Lambda$, т. е. сумма квадратов элементов *j*-го столбца в матрице нагрузок равна λ_j.

Замечание: обычно $K_{\vec{X}}$ неизвестна. Поэтому по *N* наблюдениям строят ее оценку $\widehat{K}_{\vec{X}} = \frac{1}{N-1} z^T z$, где $z -$ матрица объект центрированное свойство, и работают с ней вместо $K_{\vec{X}}$.

Пример 3.1: успеваемость в классе из 30 человек по 3-м предметам:

	$\tilde{x}_1$ Физкультура	$\tilde{x}_2$ Литература	$\tilde{x}_3$ Алгебра
1	4	5	1
2	3	2	5
…	…	…	…

30	2	5	4
$\overline{\tilde{x}_i}$	3	3,4	3

Тогда матрица объект – центрированное свойство

$$z = \begin{pmatrix} 1 & 1,6 & -2 \\ 0 & -1,4 & 2 \\ \cdots & \cdots & \cdots \\ -1 & 1,6 & 1 \end{pmatrix}, \text{ и пусть}$$

$$\widehat{K}_{\vec{X}} = \frac{1}{29} z^T z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 5 \end{pmatrix}.$$

Находя собственные значения и собственные векторы, получим

$$\Lambda = \begin{pmatrix} 6 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \; S = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{1}{\sqrt{5}} & 0 & \frac{2}{\sqrt{5}} \\ \frac{2}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \end{pmatrix}.$$

$$A = S\Lambda^{\frac{1}{2}} = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{\sqrt{6}}{\sqrt{5}} & 0 & \frac{2}{\sqrt{5}} \\ \frac{2\sqrt{6}}{\sqrt{5}} & 0 & \frac{1}{\sqrt{5}} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ -1,09 & 0 & 0,89 \\ 2,19 & 0 & 0,45 \end{pmatrix}$$

$$\begin{cases} X_1 = 0 \cdot F_1 + 1 \cdot F_2 + 0 \cdot F_3 \\ X_2 = -1,09 \cdot F_1 + 0 \cdot F_2 + 0,89 \cdot F_3 \\ X_3 = 2,19 \cdot F_1 + 0 \cdot F_2 + 0,45 \cdot F_3 \end{cases}.$$

Какие же исходные признаки X_i должны быть привлечены в выработку названия j- го главного фактора. Очевидно те, на которые F_j сильно влияет. Должны быть привлечены X_i с max $\left|a_{ij}\right|$.

Обозначим это множество исходных признаков, на которые сильно влияет F_j

$$I_j = \left\langle i: \frac{\sum_{i\in I_j} a_{ij}^2}{\sum_i a_{ij}^2} \geq 0,75 \right\rangle.$$

$K_j = \frac{\sum_{i\in I_j} a_{ij}^2}{\sum_i a_{ij}^2}$ – коэффициент информативности названия для j-го главного фактора.

Пример 3.1(продолжение):

Если взять для F_1 $I_1 = \{3\}$, то $K_1 = \frac{2,19^2}{6} = 0,8 > 0,75$.

Следовательно, название F_1 можно связать только с X_3 и назвать F_1 алгебраическими способностями.

F_1 ($c\ \lambda_1 = 6$) объясняет $\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3} = \frac{6}{8} = 75\%$ дисперсии всех переменных.

Интересно найти значения главных факторов $\vec{f}^n$ для каждого n-го объекта (с $\vec{x}^n$ – набором переменных):

$$\vec{F} = A^{-1}\vec{X}, \text{ т.е. } \quad \vec{f}^n = A^{-1}\vec{x}^n; \quad Z_{\vec{F}} = Z(A^{-1})^T.$$

Значения факторов 30 учеников:

	f_1	f_2	f_3
1	-1,03	1	0,54
2	0,99	0	-0,36
…	…	…	…
30	0,07	-1	1,9

Замечание: если с самого начала работать с оценкой корреляционной матрицы $\widehat{R}_{\vec{X}}$, то собственные значения λ_i будут другие. Ясно, $\sum_i \lambda_i = r$,

и a_{ij} имеет смысл коэффициента корреляции *i*-го стандартизированного признака с *j*-м фактором. Работать с корреляционной матрицей следует, если переменные имеют различные единицы измерения, шкалы.

Статистическая проверка надежности решения методом главных компонент

Вначале следовало бы проверить гипотезу значимости матрицы $\widehat{R}_{\vec{X}}$ (ее отличия от *I*). Если все исходные признаки окажутся попарно некоррелированными, то все λ_i будут около единицы, главные факторы станут похожи на исходные признаки, и смысла в методе главных компонент не будет. Это проверка по критерию Уилкса – Бартлетта.

Пример 3.1(продолжение):

$$\gamma = -(30 - \frac{(2 \cdot 3 + 11)}{6}) \ln \det \widehat{R}_{\vec{X}},$$

$$\gamma = -(30 - \frac{17}{6}) \ln 0,6 \; = 13,9.$$

SL(γ) = 0,003 – данные значимы против независимости исходных признаков.

3.2. Факторный анализ

Метод возник в исследованиях по психологии в 1900-х годах.

Рассмотрим успеваемость произвольно взятого ученика по *r* предметам:

$$\vec{X} = \begin{pmatrix} X_1 \\ \dots \\ X_r \end{pmatrix}.$$

Предположим, что успеваемость по этим предметам определяется небольшим числом r' скрытых факторов (способностей): F_1 – общая одарённость, F_2 – гуманитарная одарённость и т. п., так что

$$X_i = \sum_{j=1}^{r'} q_{ij} F_j + U_i, \quad i = 1, \dots, r, \qquad (18)$$

где q_{ij} – нагрузка (вклад) *j*-го фактора на *i*-й признак (отметку). Причём считаем, что q_{ij} – одинаковые для всех учеников и связаны с сутью X_i и F_j, а разные успехи определяются разными значениями F_j у разных учеников;

U_i – случайная ошибка (либо измерения отметки *i*, либо специфичность *i*-й переменной, т. е. не описываемые этими факторами вклады).

Задачи факторного анализа:

1) выявить и интерпретировать латентные (скрытые) общие факторы (т. е. оценить число r' и нагрузки q_{ij}).

2) найти оценки этих факторов для отдельных индивидуумов.

Предположения:

1. X_i – центрованы, $MX_i = 0$.

2. Вектор ошибок $\vec{U} = (U_1, \dots, U_r)^T$ не зависит от факторов F_j, состоит из взаимно независимых компонент, подчиняющихся *r*-мерному нормальному распределению $N(\vec{0}, K_{\vec{U}})$; $K_{\vec{U}} = M\vec{U}\vec{U}^T = V$ диагональная.

3. Факторы F_j – некоррелированные случайные величины с $MF_j = 0$. $DF_j = 1$, это фактически выбор масштаба, т. е. $M\vec{F}\vec{F}^T = I$.

Запишем (18) в векторной форме: $\vec{X} = Q\vec{F} + \vec{U}$. Найдем ковариационную матрицу вектора $\vec{X}$.

$$K_{\vec{X}} = M\vec{X}\vec{X}^T = M(Q\vec{F} + \vec{U})(Q\vec{F} + \vec{U})^T = M(Q\vec{F} + \vec{U})(\vec{F}^T Q^T + U^T) =$$

$$= Q(M\vec{F}\vec{F}^T)Q^T + M\vec{U}\vec{U}^T,$$

$$K_{\vec{X}} = QQ^T + V. \qquad (19)$$

Отсюда $DX_i = (K_{\vec{X}})_{ii} = \sum_{j=1}^{r'} q_{ij}(q^T)_{ji} + v_{ii} = \sum_{j=1}^{r'} q_{ij}^{\ 2} + v_{ii} = h_i^{\ 2} + v_{ii}$,

$h_i^{\ 2}$ – общность (communality), часть дисперсии, обусловленная общими факторами. v_{ii} – характерность X_i переменной.

Таким образом, если известна ковариационная матрица переменных, то нахождение нагрузок факторов на признаки – это вопрос решения уравнения (19) относительно Q и V. Это решение не единственно, и для его определения используют различные подходы:

1. Главные компоненты. Считаем собственные числа упорядоченными в порядке убывания.

По свойству главных компонент $A^T A = \begin{pmatrix} \lambda_1 \ldots\ldots\ldots\ldots\ldots 0 \\ 0 \ldots \lambda_2 \ldots\ldots\ldots\ldots 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0 \ldots\ldots\ldots\ldots \lambda_{r'} \ldots 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0 \ldots\ldots\ldots\ldots\ldots\ldots \lambda_r \end{pmatrix}$.

Матрицу A усекают, оставляя первые r' столбцов. Так получается матрица Q и

$$Q^T Q = \begin{pmatrix} \lambda_1 \ldots\ldots\ldots\ldots\ldots 0 \\ 0 \ldots \lambda_2 \ldots\ldots\ldots\ldots 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ 0 \ldots\ldots\ldots\ldots\ldots\ldots \lambda_{r'} \end{pmatrix}.$$

2. Редуцированная матрица.

По теореме Терстоуна $K_{\vec{X}} = AA^T$, и решение этого уравнения легко находится через собственные векторы. Уравнение (19) можно переписать в аналогичном

виде. $K_{\vec{X}} - V = QQ^T$, то есть по диагонали матрицы слева нужно подставить общности, которые можно оценить так:

a) $h_i^2 = \max_{j \neq i} \left| (K_{\vec{X}})_{ij} \right|$,

b) h_i^2 равно множественному коэффициенту корреляции X_i относительно остальных X_j.

Замечание: обычно ковариационная матрица неизвестна, поэтому ее заменяют оценкой, полученной по *N* наблюдениям.

$$\hat{K}_{\vec{X}} = \frac{1}{N-1} z^T z .$$

Оценка значений факторов в каждом наблюдении

Метод Бартлетта: считая нагрузки универсальными и уже найденными величинами, полагаем, что показатели отдельного объекта (наблюдения) определяются значениями его факторов. И их (его факторы) можно найти путем регрессионного анализа.

$$x_i^n = \sum_{j=1}^{r'} f_j^n q_{ij} + u_i, \ i = 1,...,r .$$

Применяя ОМНК, получаем решение:

$$\hat{\vec{f}}^n_{ОМНК} = \left[Q^T V^{-1} Q \right]^{-1} Q^T V^{-1} \vec{x}^n .$$

Практически матрицы Q и *V* в этом выражении заменяют оценочными значениями, найденными раньше.

Пример 3.2: по данным примера 3.1 на основе корреляционной матрицы провести факторный анализ, выделив один важнейший скрытый фактор.

Решение в MathCAD.

$ORIGIN := 1$

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -\frac{2}{\sqrt{10}} \\ 0 & -\frac{2}{\sqrt{10}} & 1 \end{pmatrix}$$ – оценка корреляционной матрицы.

1. Главные компоненты.

$C := eigenvals(R)$, $C^T = (1 \quad 0{,}368 \quad 1{,}632)$,

$CS := reverse(sort(C))$,

$CS^T = (1{,}632 \quad 1 \quad 0{,}368)$ – упорядоченные собственные числа.

$i := 1, \ldots, 3.$

$V_i := eigenvec(R, CS_i)$,

$S := augment(V_1, V_2, V_3)$,

$$S = \begin{pmatrix} 0 & 1 & 0 \\ -0{,}707 & 0 & 0{,}707 \\ 0{,}707 & 0 & 0{,}707 \end{pmatrix}$$ – матрица из собственных векторов.

$\Lambda sr := diag(\sqrt{CS})$, $A := S\Lambda sr$,

$$A = \begin{pmatrix} 0 & 1 & 0 \\ -0{,}903 & 0 & 0{,}429 \\ 0{,}903 & 0 & 0{,}429 \end{pmatrix}$$ – матрица нагрузок метода главных компонент.

Усекаем ее, оставляя один важнейший фактор:

$Q := submatrix(A, 1, 3, 1, 1)$,

$$Q = \begin{pmatrix} 0 \\ -0{,}903 \\ 0{,}903 \end{pmatrix}$$ – матрица нагрузок факторного анализа. Единственный фактор следует интерпретировать как алгебраические – анти литературные способности.

$V := R - QQ^T$,

$V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0{,}184 & 0{,}184 \\ 0 & 0{,}184 & 0{,}184 \end{pmatrix}$ – по диагонали – характерности.

$Factor\ Scores\ Coeff := (Q^T V^{-1} Q)^{-1} Q^T V^{-1}$,

$Factor\ Scores\ Coeff = (0 \quad -0{,}553 \quad 0{,}553)$ – коэффициенты для подсчета значения скрытого фактора для объекта по его переменным. В частности, для первого ученика

$f1 := Factor\ Scores\ Coeff \begin{pmatrix} 1 \\ 1{,}6 \\ -2 \end{pmatrix}$, $f1 = -1{,}99$. Большое отрицательное значение фактора алгебраических – анти литературных способностей.

2. Редуцированная матрица.

$Rr := \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{2}{\sqrt{10}} & -\frac{2}{\sqrt{10}} \\ 0 & -\frac{2}{\sqrt{10}} & \frac{2}{\sqrt{10}} \end{pmatrix}$. По диагонали общности $h_i^2 = \max_{j \neq i} \left| R_{ij} \right|$.

Далее решаем уравнение $Rr = QQ^T$.

$$Q = \begin{pmatrix} 0 & 0 & 0 \\ -0.795 & 0 & 0 \\ 0.795 & 0 & 0 \end{pmatrix}.$$

Это ведет к подобным $Factor\ Scores\ Coeff = (0 \quad -0{,}629 \quad 0{,}629)$.

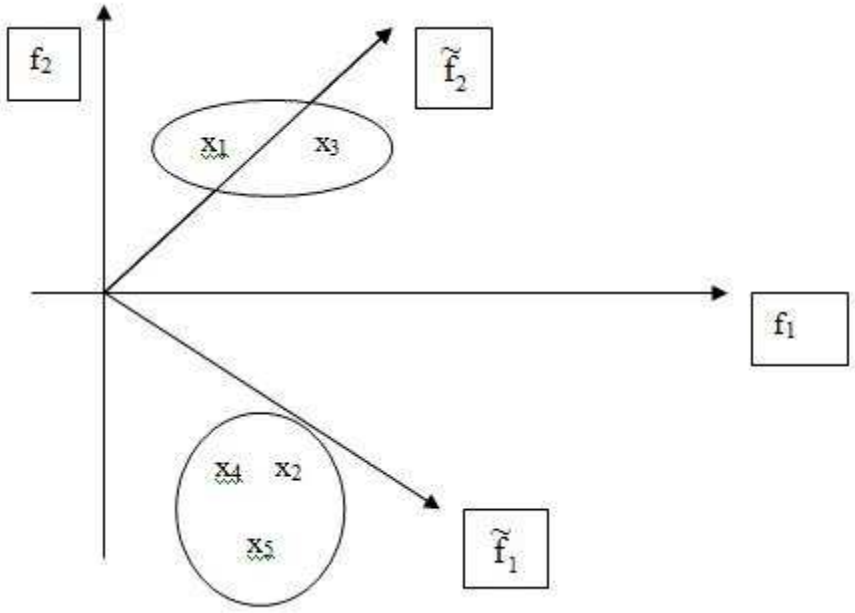
Вращение факторов

Как уже отмечалось, решение уравнения (19) не единственно. В частности, если Q решение, то и $\tilde{Q} = QB^{-1}$, где B ортогональная матрица – тоже решение. Оно соответствует другим факторам, связанным со старыми формулой

$$\tilde{\vec{F}} = B\vec{F} \quad . \qquad (20)$$

Так как B – ортогональная матрица, (20) означает поворот системы координат. Можно подобрать B так, чтобы в $\tilde{Q}$ нагрузки факторов были большими для одних признаков и близкими к нулю для других (процедура ВАРИМАКС). Это облегчит интерпретацию новых факторов.

Пример 3.3: пусть для пяти признаков $x_1 - x_5$ получены нагрузки (матрица Q) как на рисунке.



$q_{32} = 0.8$; $q_{22} = -0.8$. f_2 – способность к x_3 и анти- к x_2. f_1 – способность к x_3 и x_2. Ясно, что целесообразно повернуть систему координат: $\tilde{f}_1, \tilde{f}_2$. Тогда интерпретация облегчится, $\tilde{f}_1$ имеет большую нагрузку на признак x_2 и близкую к нулю на признак x_3. $\tilde{f}_2$ – наоборот. То есть $\tilde{f}_1$ – способность к x_2, $\tilde{f}_2$ – способность к x_3. После вращения можно оценить значения новых факторов для всех наблюдений, как и прежде.

3.3. Упражнения и лабораторный практикум

Упражнение 3.1.

Доказать формулы (13) – (17).

Упражнение 3.2.

Применить критерий Уилкса – Бартлетта в примере 3.1.

Упражнение 3.3.

Применить анализ главных компонент к данным примера 1.1.

Statistica: Statistics, Multivariate Exploratory Techniques, Principal Components & Classification Analysis, все переменные выбрать в первый столбец, Advanced проверить, что анализ базируется на корреляционной матрице и делится на (N-1) , OK.

Eigenvalues смотрим собственные числа, Scree plot –осыпь.

Factor coordinates of the variables – матрица нагрузок. Plot var. factor coordinates, 2D – соответствующий график, factor coordinates of cases и график – значения главных компонент (ненормированных!) для наблюдений.

Variables Tab, Contributions of variables – доли (относительно соответствующего лямбда) квадратов нагрузок – удобно набрать в сумме
>= 0,75, указать коэффициент информативности; сформулировать соответствующую интерпретацию (названия) главных компонент (с оглядкой на знаки!).

Cases Tab, Factor score coefficients – матрица (A^T)-1 –для расчёта нормированных факторов для каждого наблюдения. Factor scores – значения этих факторов (проверить, скопировав матрицу нагрузок в MathCAD, исходные данные стандартизировать!). Охарактеризовать нескольких учеников в терминах факторов, с подтверждением в терминах исходных переменных. Построить 3D график наблюдений в пространстве факторов.

Descriptives Tab, Correlation matrix Inverse. Видим логарифм определителя корреляционной матрицы для теста Уилкса – Бартлетта. Сделать выводы.

Упражнение 3.4.

Повторить в MathCAD результаты метода главных компонент, полученные с применением Statistica в упражнении 3.3.

Упражнение 3.5.

Проанализировать методом главных компонент свои данные (4 – 6 количественных переменных в 15 – 20 наблюдениях). Например: вкус (желание) к продукту питания (по 5-балльной шкале). X_1 – у девочек до 15 лет, X_2 – у женщин после 20 лет, X_3 , X_4 – аналогично у мужчин. Наблюдения (продукты) – кофе, шоколад, сыр и т. д.

Упражнение 3.6.

Доказать формулу (20).

Упражнение 3.7.

Повторить пример 3.2 в Statistica, построив файл матричного формата для корреляционной матрицы (Statistica: Help, Working with Spreadsheets, Understanding Spreadsheets, Matrix Spreadsheets, Matrix File Format).

	VAR1	VAR2	VAR3
VAR1	1,000	0,000	0,000
VAR2	0,000	1,000	-0,632
VAR3	0,000	-0,632	1,000
MEANS	3,000	4,000	3,000
ST.DEV.	2,000	2,000	2,000
NO.CASES	30,000		
MATRIX	1,000		

Прокрутив вверху Max. number of factors, выбрать = 1. Method: Principal components, Maximum likelihood factors.

Упражнение 3.8.

Провести факторный анализ данных примера 1.1. Выделить 2 фактора. Обсудить communalities. Использовать вращение факторов для облегчения их интерпретации. Дать факторам названия с указанием коэффициентов информативности. Указать оценки факторов для всех объектов. Сравнить с результатами анализа главных компонент.

Упражнение 3.9.

В результате наблюдения за экологической обстановкой в семи городах с различным уровнем техногенной нагрузки на окружающую среду получены данные:

	X_1	X_2	X_3	X_4
1	0,14	0,005	1,6	0,02
2	0,1	0,004	1,2	0,04
3	0,25	0,005	2,4	0,05
4	0,27	0,01	1,7	0,04
5	0,22	0,07	3	0,08
6	0,16	0,012	1,8	0,06
7	0,21	0,03	1,1	0,04

X_1 – средняя концентрация загрязняющих веществ в атмосферном воздухе (мг/м3), пыль,

X_2 – то же, сернистый ангидрит,

X_3 – то же, окись углерода,

X_4 – то же, двуокись азота.

Провести факторный анализ загрязнения воздуха в Statistica.

Упражнение 3.10.

Используя те же данные, что и в упражнении 3.5, провести факторный анализ, применив вращение факторов для облегчения их интерпретации. Указать оценки факторов для всех объектов.

4. Многомерное шкалирование

Иногда исходная информация об объектах 1, 2, …, *N* бывает задана в форме матрицы их попарных сравнений

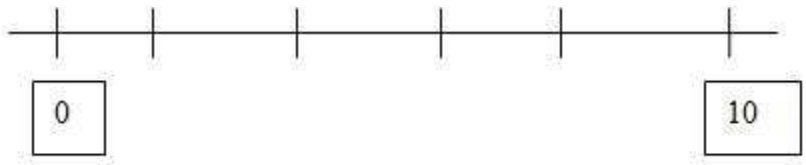
$$\gamma = \begin{pmatrix} \gamma^{11} & \dots & \gamma^{1N} \\ \dots\dots\dots\dots & & \\ \gamma^{N1} & \dots & \gamma^{NN} \end{pmatrix}.$$

γ^{mn} могут иметь смысл:

1. Евклидово расстояние между объектами *m* и *n* в некотором *r*-мерном признаковом пространстве:

$$\gamma^{mn} = \sqrt{\sum_{k=1}^{r} (x_k{}^m - x_k{}^n)^2} = \sqrt{(\vec{x}^m - \vec{x}^n)^T (\vec{x}^m - \vec{x}^n)}. \qquad (21)$$

2. Рейтинговые оценки различий объектов. Экспертам предлагается шкала с некоторым числом делений (до 7 – 10), позволяющая оценивать каждую пару объектов по степени их сходства; 0 – полное сходство, 10 – абсолютное различие.



Пример 4.1: различия 5 государств (Армения, Беларусь, Россия, Таджикистан, Литва) с учетом их экономического и политического положения, полученные экспертами:

$$\gamma = \begin{pmatrix} & А & Б & Р & Т & Л \\ А & 0 & 10 & 9 & 3 & 7 \\ Б & 10 & 0 & 1 & 5 & 2 \\ Р & 9 & 1 & 0 & 4 & 6 \\ Т & 3 & 5 & 4 & 0 & 8 \\ Л & 7 & 2 & 6 & 8 & 0 \end{pmatrix}.$$

Задача многомерного шкалирования: по матрице γ восстановить неизвестную размерность признакового пространства и приписать каждому *n*-у объекту вектор $\vec{x}^n$ этих признаков таким образом, чтобы вычисленные по формуле (21) попарные евклидовы расстояния по возможности совпали с исходной матрицей γ. (То есть задача сугубо обратная).

4.1. Метрическое многомерное шкалирование

Разработано У. Торгерсоном в 1950-е годы.

Будем считать векторы $\vec{x}^n$ центрованными. Очевидно, расстояния между векторами центрованными $\vec{x}^n$ и не центрованными $\tilde{\vec{x}}^n$ одинаковы.

$$\gamma^{mn} = \tilde{\gamma}^{mn}. \qquad (22)$$

Так как расстояния между векторами не меняются при ортогональном преобразовании, то $\vec{x}^n$ определены с точностью до поворота системы координат.

Введем матрицу

$B_{[N\times N]} = \left(b^{mn}\right)$, где $b^{mn} = \vec{x}^{mT} \cdot \vec{x}^n$.

Она связана с матрицей γ:

$$b^{mn} = \frac{1}{2}\left(-\gamma^{mn2} + \frac{1}{N}\sum_m \gamma^{mn2} + \frac{1}{N}\sum_n \gamma^{mn2} - \frac{1}{N^2}\sum_m\sum_n \gamma^{mn2} \right). \qquad (23)$$

Если $Z_{[N\times r]} = \begin{pmatrix} \vec{x}^{1T} \\ ... \\ \vec{x}^{NT} \end{pmatrix}$ матрица объект – центрированное свойство, то легко проверить, что

$$B = ZZ^T. \qquad (24)$$

$\hat{K}_{\vec{x}} = \frac{1}{N-1} Z^T Z$, обозначим $K' = Z^T Z$. Из связи матриц K' и B вытекают следствия:

1) B неотрицательно определена,

2) rank B = rank K' = r,

3) ненулевые (первые *r*) собственных чисел матрицы *B* совпадают с собственными числами матрицы K'.

Чтобы найти Z в (24), используем сингулярное разложение:

$B = ZZ^T = S_B \Lambda S_B{}^T$. Находим $Z = S_B{}^{\langle 1 \div r \rangle} \Lambda_{[r \times r]}^{\frac{1}{2}}$, или

$$x_k{}^n = e_n{}^k \sqrt{\lambda_k}, \qquad (25)$$
$$\left(n = 1,...,N; \quad k = 1,...,r\right)$$

где $\vec{e}^k$ – *k*-й собственный вектор матрицы *B*.

Замечание: если мы хотим представить имеющиеся данные в пространстве заданной (< *r*) размерности r', то можно показать, что, взяв только первые r' координат в формуле (25), мы обеспечим максимальное приближение их геометрической структуры. При $r' = 2$ или 3 возможно наглядное представление.

Пример 4.2: сведения об обороте капитала, прибыли, количестве работников некоторых компаний в 1993 году:

Компания	Оборот капитала, млрд. $.	Прибыль, млн. $.	Количество работников, тыс. чел.
МАН	11,8	429,6	63,4
СЕБ	1,5	88,8	10,1
Даниска	2	124,6	11,5
Нокиа	3,1	-17,1	26,8

Вначале нормируем исходные данные по формуле $x_k'^n = \frac{x_k^n}{\overline{x}_k}$, затем считаем по формулам (21) – (25):

$$\gamma = \begin{pmatrix} 0 & 3,66 & 3,43 & 3,67 \\ 3,66 & 0 & 0,258 & 0,967 \\ 3,43 & 0,258 & 0 & 1,08 \\ 3,67 & 0,967 & 1,08 & 0 \end{pmatrix},$$

$$B = \begin{pmatrix} 7,105 & -2,62 & -1,978 & -2,49 \\ -2,62 & 1,049 & ..... & \\ ..... & & & \end{pmatrix},$$

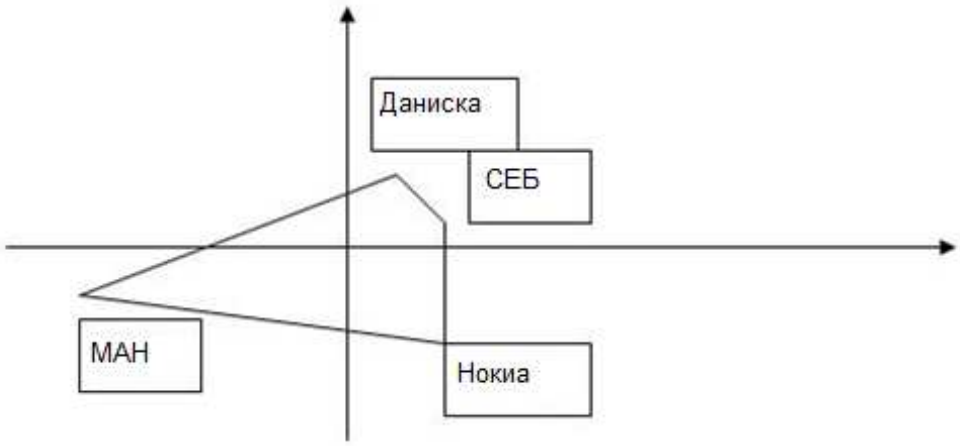
$$\lambda_1 = 9,506 \quad \lambda_2 = 0,694 \quad \lambda_3 = 0 \quad \lambda_4 = 0 \quad \vec{e}^1\sqrt{\lambda_1} = \begin{pmatrix} -2,67 \\ 0,98 \\ 0,74 \\ 0,94 \end{pmatrix},$$

$$\vec{e}^2\sqrt{\lambda_2} = \begin{pmatrix} -0,02 \\ 0,295 \\ 0,395 \\ -0,671 \end{pmatrix}.$$

Таким образом, координаты объектов (стимулов) в двумерном шкальном пространстве:

$$\begin{Bmatrix} МАН \\ СЕБ \\ Даниска \\ Нокиа \end{Bmatrix} \begin{pmatrix} -2,67 & -0,02 \\ 0,98 & 0,259 \\ 0,74 & 0,395 \\ 0,94 & -0,671 \end{pmatrix}.$$

$$\vec{x}^1 = \begin{pmatrix} -2,67 \\ -0,02 \end{pmatrix} \text{ и т. д.}$$



Интерпретация шкал. Судя по расположению фирм, x_1 можно назвать осью размера (малости) фирмы, x_2 – эффективностью деятельности.

Кстати, расстояния по новым координатам

$\gamma_{нов}^{mn} = \sqrt{(x_1^m - x_1^n)^2 + (x_2^m - x_2^n)^2} = \gamma^{mn}$, то есть восстанавливаются точно.

4.2. Неметрическое многомерное шкалирование

Применяется для обработки экспертных зачастую ранговых (порядковых) данных.

Пример (продолжение 4.1): из пяти государств можно образовать

$$C_5^{\,2} = \frac{5!}{2!\cdot 3!} = 10$$

сочетаний по два. То есть имеется 10 расстояний (отличий и т.п.), измеренных неизвестно в каких единицах, и можно их только упорядочить:

$$(\text{Россия}, \text{Беларусь}) < (\text{Беларусь}, \text{Литва}) < (\text{Таджикистан}, \text{Армения}) < \ldots$$

$$\ldots < (\text{Армения}, \text{Беларусь}).$$

Ранги 1, 2, 3, …, 10.

Основная проблема в том, чтобы расстояния в шкальном пространстве между объектами монотонно соответствовали рангам:

$$\gamma_{ранг}^{mn} < \gamma_{ранг}^{kl} \Leftrightarrow \gamma_{\vec{x}}{}^{mn} < \gamma_{\vec{x}}{}^{kl}.$$

Обычно используется итерационный алгоритм:

1) по $\gamma_{ранг}$ находят первоначальную оценку шкальных координат (стартовую конфигурацию) $\gamma_{ранг} \to \{\vec{x}^n\}$. Иногда ее выбирают случайной;

2) находят расстояния $\gamma_{\vec{x}}{}^{mn}$ и сопоставляют их с $\gamma_{ранг}^{mn}$. Если нарушена монотонность, подправляют;

3) проверяется хорошая воспроизводимость в смысле стресс-критерия,

$S = \dfrac{\sum_{mn} (\gamma_{\vec{x}}{}^{mn} - \gamma_{ранг}^{mn})^2}{\sum_{mn} (\gamma_{\vec{x}}{}^{mn})^2} \to \min$. Если необходимый минимум достигнут, то делается переход к п. 5, иначе – к п. 4;

4) находят новые координаты (наискорейший спуск с определенным шагом), переход к п. 2;

5) финальная конфигурация в шкальном пространстве $\left\{\vec{x}^n\right\}$.

Если шкалировать в одномерное пространство, то финальная конфигурация сразу дает веса объектов – понятий (после сдвига и нормировки).

4.3. Шкалирование индивидуальных различий экспертов

Иногда целью исследования является установление особенностей (индивидуальных различий) экспертов, доставляющих матрицы попарных сравнений объектов.

Обозначим: γ_s^{mn} – матрица расстояний между всеми парами объектов, представленная *s*-м экспертом. Предполагается, что

$$\gamma_s^{mn} = \sqrt{\sum_{k=1}^{r} w_k^s (x_k{}^m - x_k{}^n)^2} = \sqrt{(\vec{x}^m - \vec{x}^n)^T W^s (\vec{x}^m - \vec{x}^n)},$$

где $W^s = \mathrm{diag}(w_1^s, w_2^s, ..., w_r^s)$ – матрица весов, которые невольно использует (вносит) *s*-й эксперт при оценивании различий между объектами по разным показателям. Замечательно, что мы еще не знаем (и эксперт сам не осознает), сколько этих показателей, каков их смысл, а теперь еще пытаемся выделить предпочтения по ним!

Следовательно, предполагается, что есть общая объективная конфигурация $\{\vec{x}^n\} \Rightarrow Z$, а различия между γ_s определяются различием матриц W^s.

Процедура:

1) по γ_s вычисляется B_s (23),
2) находятся средняя матрица скалярных произведений $\overline{B} = \frac{1}{S}\sum_{s=1}^{S} B_s$ и стартовая конфигурация Z_0,
3) решается оптимизационная задача

   $\sum_{s=1}^{S} \left\| B_s - ZW^s Z^T \right\| \to \min_{Z, W^s}$ для нахождения финальной конфигурации Z и весов W^s.

4.4. Анализ предпочтений

Здесь каждый из S экспертов представляет ранжировку N объектов по предпочтению относительно какой-либо цели. r_s^n – ранг (место, 1 – наилучшее) n-го объекта у s-го эксперта.

Модель (взвешенная евклидова):

$$r_s^n = a_0 + \sum_{k=1}^{r} w_k^s (x_k^n - o_k^s)^2 \ , \qquad (26)$$

o_k^s – k-я координата "идеальной" точки для s-го эксперта, и чем ближе к ней показатели n-го объекта, тем объект предпочтительнее; w_k^s – веса, характеризующие важность координаты k для эксперта s; a_0 – некоторая константа.

Анализ делится на две части: внутренний и внешний. Во внутреннем анализе решается задача оценки координат $\left\{\vec{x}^n\right\}$ объектов обычными методами многомерного шкалирования. Например, можно по r_s^n построить $\gamma_s^{nm} = \left| r_s^n - r_s^m \right|$, затем $\gamma_s \to \overline{\gamma} \to \overline{B} \to$ – финальную конфигурацию.

Во внешнем анализе, считая $\left\{\vec{x}^n\right\}$ известными, проводим оценку параметров модели (26) каждого эксперта: идеальную точку и веса. Для этого используем регрессионный анализ:

$$\breve{r}_s^n = a_0 + \sum_{k=1}^{r} (w_k^s x_k^{n2} - 2 w_k^s x_k^n o_k^s + w_k^s o_k^{s2}) =$$

$$= \beta_0^s + \sum_{k=1}^{r} \beta_k^s x_k^n + \sum_{k=1}^{r} \mu_k^s x_k^{n2} \ .$$

Найдем оценки $\hat{\beta}^s, \hat{\mu}^s \to \hat{W}^s, \hat{\vec{o}}^s, \hat{a}_0$ (в частности, $o_k^s = \dfrac{\beta_k^s}{-2 w_k^s}$ – координаты идеальной точки s-го эксперта).

Пример 4.3: 2 эксперта ранжировали 6 областей Беларуси по условиям организации сельскохозяйственного производства:

n	Область	r_1^n	r_2^n
1	Брестская	2	1
2	Витебская	4	3
3	Гомельская	6	6
4	Гродненская	3	4
5	Минская	1	2
6	Могилевская	5	5

$$\gamma_1 = \begin{pmatrix} 0 & & & & & \\ 2 & 0 & & & & \\ 4 & 2 & 0 & & & \\ 1 & 1 & 3 & 0 & & \\ 1 & 3 & 5 & 2 & 0 & \\ 3 & 1 & 1 & 2 & 4 & 0 \end{pmatrix}, \quad \gamma_2 = \begin{pmatrix} 0 & & & & & \\ 2 & 0 & & & & \\ 5 & 3 & 0 & & & \\ 3 & 1 & 2 & 0 & & \\ 1 & 1 & 4 & 2 & 0 & \\ 4 & 2 & 1 & 1 & 3 & 0 \end{pmatrix},$$

$\gamma_\Sigma = \gamma_1 + \gamma_2$.

Или загружая ранги в Statistica, Cluster Analysis, metrics Manhattan (City block): $\gamma_\Sigma^{nm} = \sum_{s=1}^{2} \left| r_s^n - r_s^m \right|$, получим нужную матрицу γ_Σ, сохраним в файле с расширением .smx.

По γ_Σ, считая это матрицей расстояний, получим финальную конфигурацию:

$\begin{pmatrix} -0,97 & 0,205 \\ -0,04 & -0,507 \\ \dots & \dots \\ \dots & \dots \end{pmatrix}$. На этом внутренний анализ завершен.

Внешний анализ для первого эксперта. В Statistica выбираем r_1 зависимой переменной, x_1, x_2, x_1^2, x_2^2 – независимыми; получим:

$$\hat{\beta}_0^1 = 6,3, \hat{\beta}_1^1 = 1,92, \hat{\beta}_2^1 = 3,7, \hat{\mu}_1^1 = -3,3, \hat{\mu}_2^1 = -1,4.$$

Таким образом, $\hat{\hat{o}}^1 = \begin{pmatrix} 0,29 \\ 1,32 \end{pmatrix}$, $\hat{W}^1 = \begin{pmatrix} -3,3 & 0 \\ 0 & -1,4 \end{pmatrix}$. Так как веса отрицательны, точка $\hat{\hat{o}}^1$ фактически является «антиидеальной» для первого эксперта.

Для второго эксперта $\hat{\bar{\delta}}^2 = \begin{pmatrix} -0,38 \\ 0,6 \end{pmatrix}$, $\hat{W}^2 = \begin{pmatrix} 2,4 & 0 \\ 0 & 1,85 \end{pmatrix}$. Ясно, почему он считает лучшей Брестскую область.

Если использовать вектор средних рангов (в случае согласованности экспертов), то результат будет объективным (идеальная точка), к которому следует стремиться.

4.5. Упражнения и лабораторный практикум

Упражнение 4.1.

Доказать формулы (22) – (23).

Упражнение 4.2.

Доказать формулы (24) – (25).

Упражнение 4.3.

Получить результаты примера 4.2 (фирмы МАН, СЕБ, Даниска, Нокиа).

1. Предварительно построить файл матричных данных для расстояний между объектами:
   а. нормируем raw – данные, поделив каждое данное на выборочное среднее соответствующей переменной (пометить столбец, правый клик – Statistics of Block Data - Block Columns – Means. После последнего case появится строка со средним. Пометить столбец, правый клик – Var Spec. = V?/ (вставить соотв. среднее).
   б. с помощью процедуры Cluster Analysis –Joining – Variables – Var1-3 – Cluster – Cases (rows) – Distance measure – Euclid – OK – Advanced – Matrix. File – Save As ... сохраним матрицу расстояний.
2. Используя сохранённую матрицу расстояний, провести многомерное шкалирование. Multivariate Exploratory Techniques. Multidimensional Scaling. Variables: 4. Number of dimensions: 2. Summary: Final configuration. Graph final configuration, 2D.

Упражнение 4.4.

Провести неметрическое шкалирование данных примера 4.1.

Матрицу расстояний проще всего ввести в нужном формате, редактируя: – Matrix. File – Save As ... сохраним матрицу расстояний (см. 1. б). Интерпретировать шкалы (возможно лучше с поворотом).

Упражнение 4.5.

Повторить в MathCAD пример 4.2. Убедиться в точном восстановлении расстояний по финальной конфигурации.

Упражнение 4.6.

2 эксперта сравнивали 4-х людей. 1-й высокий и полный, 2-й высокий худой, 3-й низкорослый полный, 4-й низкорослый худой. Матрицы попарных сравнений у экспертов

$$\gamma_1 = \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 5 & 7 & 0 & \\ 6 & 4 & 2 & 0 \end{pmatrix}, \qquad \gamma_2 = \begin{pmatrix} 0 & & & \\ 5 & 0 & & \\ 1 & 6 & 0 & \\ 7 & 2 & 5 & 0 \end{pmatrix}.$$

Выполнить в MathCAD шкалирование индивидуальных различий экспертов.

По средней матрице скалярных произведений получить стартовую конфигурацию, с помощью функции Minimize() получить финальную конфигурацию, интерпретировать шкалы и найти весовые коэффициенты, характеризующие индивидуальные различия экспертов.

Упражнение 4.7.

Повторить пример 4.3.

1. Внутренний анализ. В процедуре Cluster Analysis, Joining, Distance measure – Manhattan (City-block) получить среднюю матрицу расстояний и сохранить ее. Многомерным шкалированием получить финальную конфигурацию. Дать интерпретацию шкал.
2. Внешний анализ в рамках взвешенной евклидовой модели провести для каждого эксперта. Advanced Linear/Nonlinear Models, Fixed Nonlinear Regression. Найти оценки коэффициентов параболической функции двух полученных переменных. Вычислить оценки координат идеальной точки и весов и с их помощью объяснить ранжировку каждого эксперта. Как правильнее было бы назвать точку у 1-го эксперта?

Считая мнения экспертов согласованными, найти "объективно" идеальную точку и веса.

5. Расчётно-графическая работа

Индивидуальная расчётно-графическая работа состоит из 6 заданий. Отчёт по каждому заданию высылается преподавателю по e-mail отдельным файлом к объявленному сроку. В теме сообщения и имени файла обязательно указать номер группы, свою фамилию и номер задания. Формат файла – MS Word. Подобранные собственные исходные данные должны быть приведены полностью в виде доступном для электронной обработки (с целью проверки). Пункты отчёта должны содержать ответы на все вопросы задания (наименование и значение). Расчёты по собственным данным должны быть подтверждены таблицами из Statistica или страницами из MathCAD, которые вставляются в виде рисунков.

Задание № 1.

Для 12 промышленных предприятий собраны данные (см. свой вариант). Первая строка – производительность труда (т/ч), вторая – уровень травматизма (число случаев на 1000 работников), третья – средний возраст работников (лет), четвёртая – энерговооружённость (КВт/100 работающих).

1. Оценить коэффициент корреляции производительности труда и уровня травматизма, значимость его отличия от нуля (*sl*) (о1, о2).
2. Оценить частный коэффициент корреляции производительности труда и уровня травматизма, значимость его отличия от нуля при исключении влияния энерговооружённости труда (о3, о4). Объяснить механизм ложной корреляции, проиллюстрировать возникновение ложной корреляции диаграммой, подобной примеру 1.3 из пособия, сделать выводы.
3. Оценить частный коэффициент корреляции производительности труда и уровня травматизма, значимость его отличия от нуля при исключении влияния среднего возраста и энерговооружённости труда (о5, о6).

4. Оценить множественный коэффициент корреляции производительности труда со всеми остальными переменными (о7). Подтвердить результат с помощью регрессионного анализа.
5. Проверить критерием Уилкса – Бартлетта гипотезу о попарной независимости производительности труда, уровня травматизма, среднего возраста. Указать *sl* (о8).

   Форма ответа: о1, о2, о3, о4, о5, о6, о7, о8.

Вариант 1											
132	132	133	133	135	132	134	132	131	131	135	132
64	66	68	66	68	65	69	67	67	67	69	66
64	51	35	55	36	53	26	39	51	58	37	55
300	420	650	510	790	410	730	480	360	390	890	370
Вариант 2											
106	108	106	107	106	105	105	107	107	105	106	107
62	62	63	63	64	61	63	62	64	61	63	63
33	51	43	38	56	49	44	54	39	62	51	32
580	750	660	740	640	390	560	610	750	400	600	720
Вариант 3											
70	69	68	70	69	71	70	71	67	69	69	68
53	51	49	50	52	51	52	51	51	51	50	51
43	44	55	33	33	38	30	27	61	58	38	45
800	600	320	640	620	810	700	810	370	490	470	390
Вариант 4											
16	15	16	17	16	16	14	16	15	17	16	16
57	56	57	57	58	57	56	56	55	55	57	55
43	41	41	29	49	53	58	37	35	51	46	31
680	570	660	810	680	640	410	630	450	660	650	600
Вариант 5											
112	110	111	114	112	112	113	112	113	113	111	110
7	5	5	7	8	5	5	7	5	5	4	6
48	40	37	34	38	53	44	52	31	39	39	35
690	360	440	850	700	600	600	610	710	660	300	360

Вариант 6											
83	83	81	84	81	85	84	82	81	82	83	81
37	36	37	37	37	37	37	36	37	36	37	35
45	48	44	25	49	41	32	30	34	57	52	55
650	630	490	790	470	840	750	520	440	440	630	330
Вариант 7											
27	25	27	26	28	25	29	27	29	26	27	26
75	76	77	73	75	75	76	74	76	75	74	74
34	52	53	62	31	51	51	49	30	42	54	34
620	390	740	320	690	350	870	470	890	480	520	430
Вариант 8											
60	58	59	59	59	58	60	61	58	57	60	59
13	12	15	15	13	14	14	15	14	13	13	14
34	52	47	30	30	39	32	25	52	53	47	30
660	420	690	660	570	540	660	870	540	410	650	680
Вариант 9											
133	135	136	136	134	134	134	135	135	135	135	135
51	53	53	52	53	51	50	52	50	53	52	53
58	47	43	47	30	44	56	57	54	56	45	51
330	580	740	710	520	330	400	490	430	590	560	660
Вариант 10											
4	6	6	6	6	4	7	6	4	6	5	6
55	59	58	56	56	56	57	56	54	59	54	55
45	28	45	47	43	33	33	52	36	46	57	38
420	780	810	680	670	450	770	730	350	810	490	580
Вариант 11											
56	57	57	56	55	55	55	54	57	54	56	54
53	53	55	56	53	52	54	51	53	52	54	52
56	29	45	47	48	51	46	51	33	48	46	51
690	800	820	850	570	510	560	300	770	350	710	390
Вариант 12											
129	129	128	129	128	128	130	127	129	128	129	128
77	79	77	76	76	76	77	76	77	77	75	76
44	45	56	38	56	43	49	33	54	32	30	31

780	890	680	660	600	570	870	450	740	680	690	630
Вариант 13											
87	88	88	88	88	85	86	87	89	86	86	85
48	46	47	47	49	46	47	48	48	47	45	46
49	48	54	49	34	34	30	50	28	50	43	58
650	740	700	790	900	420	530	650	880	530	370	330
Вариант 14											
113	111	113	110	109	110	113	112	112	111	111	112
59	55	57	55	56	56	57	59	59	55	55	59
45	35	35	61	34	44	32	29	39	54	38	47
890	480	830	360	390	490	790	790	840	460	440	830
Вариант 15											
140	140	139	140	138	138	137	138	138	138	138	140
66	67	66	68	65	66	66	67	67	64	64	68
47	45	43	32	51	38	48	40	45	44	34	42
630	810	620	820	440	520	320	540	500	420	380	840
Вариант 16											
42	41	40	44	43	42	43	41	45	44	43	43
14	15	13	15	14	15	14	15	15	17	13	13
44	31	45	40	38	45	42	59	48	47	54	58
560	530	310	820	720	660	640	530	860	880	650	550
Вариант 17											
106	102	104	102	103	103	103	105	105	105	105	104
80	78	79	77	80	77	78	79	79	79	80	79
26	50	56	60	41	53	40	54	44	29	42	39
870	380	580	320	570	380	430	780	700	770	810	600
Вариант 18											
79	81	82	80	81	79	80	79	79	82	81	81
11	12	11	10	12	10	11	10	10	11	12	11
58	42	34	54	31	32	33	37	55	26	29	43
590	870	890	610	840	500	640	510	480	850	790	710
Вариант 19											
19	19	18	18	20	20	19	19	19	18	18	18
62	61	61	60	63	62	64	63	63	60	60	60

47	29	49	40	46	51	43	50	44	44	35	32
680	570	500	470	820	820	780	680	660	420	440	440
Вариант 20											
66	66	68	65	65	66	69	68	66	68	68	65
49	50	53	50	50	49	52	52	52	51	53	50
42	60	37	57	38	36	37	34	48	35	36	41
460	440	820	330	330	350	880	830	540	700	870	300
Вариант 21											
30	29	32	33	32	30	31	31	33	31	31	29
44	41	42	43	45	41	42	42	43	41	45	41
36	43	50	42	24	41	55	51	52	32	28	45
610	310	760	880	840	430	590	550	810	560	740	350
Вариант 22											
137	136	135	138	138	135	137	139	136	138	138	136
39	38	39	41	41	39	41	39	40	42	42	38
36	49	57	38	42	44	54	36	38	26	30	36
610	390	350	850	750	360	710	800	460	820	880	430
Вариант 23											
147	145	147	146	144	148	144	144	144	144	146	145
43	42	43	43	43	43	42	41	41	42	44	41
55	43	45	42	41	47	58	48	47	45	44	57
670	420	760	680	410	880	420	330	320	350	640	430
Вариант 24											
64	65	64	67	64	67	65	65	66	67	66	65
13	12	15	14	15	15	13	14	14	17	14	14
39	37	52	41	53	43	57	57	56	45	41	57
310	330	450	730	380	820	430	490	670	870	530	440
Вариант 25											
79	77	80	79	76	78	77	77	77	79	79	77
59	57	59	58	58	59	58	58	58	58	59	59
36	55	25	50	57	51	49	35	35	31	47	57
750	440	770	640	380	550	400	450	440	720	690	500

Задание № 2.

1. Подобрать данные для корреляционного анализа 3 – 4-х ординальных переменных (10 – 15 наблюдений).
2. Оценить коэффициенты корреляции Спирмена, найти значимости их отличий от нуля. Сделать выводы, дать результатам содержательное объяснение. Рассчитать среднее (арифметическое) по всем парам значение коэффициента корреляции Спирмена.
3. Оценить коэффициент конкордации Кендалла, найти значимость отличия от нуля.
4. Проверить формулу связи коэффициента конкордации и среднего коэффициента корреляции Спирмена.
5. Считая ранжировки согласованными, вывести итоговое упорядочение.

Задание № 3.

1. Подобрать данные для корреляционного анализа 3 – 4 х категоризованных переменных (10 – 15 наблюдений). Например: группа крови, знак зодиака, черты характера.
2. Получить таблицу сопряжённости какой-либо пары переменных, проверить гипотезу их независимости, найти коэффициент сопряжённости Крамера.
3. Добавить третью переменную. Исследовать зависимость пары переменных в различных категориях третьей переменной, описать суть выявленных зависимостей. Проверить значимость выводов (найти sl).
4. Проверить гипотезу независимости всех имеющихся переменных. Найти sl.

Задание № 4.

1. Подобрать данные (8 – 15 наблюдений) для дискриминантного анализа по 2 – 4 признакам в 2 – 3 класса.
2. Указать смысл классов.
3. Найти оценки центров классов, ковариационной матрицы и априорных вероятностей.
4. Классифицировать новое наблюдение, вычислив значения линейных классифицирующих функций. Дать содержательную интерпретацию, почему новое наблюдение попало в соответствующий класс.
5. Классифицировать все наблюдения (в STATISTICA: … Classification of cases), сравнить классификацию обучающих наблюдений с фактической, ещё раз подтвердить классификацию нового наблюдения.
6. Проверить гипотезу о неразличимости классов (указать sl).

Задание № 5.

1. Подобрать данные для кластерного анализа: 10 – 15 наблюдений (3 – 4 мерных).

2. Методом K-средних расклассифицировать наблюдения в 2 класса, указать состав классов, проверить гипотезу существенности различий между центрами классов по каждой переменной, для самой значимой указать sl. Сравнить это с Plot of means.

3. Стандартизировать данные и привести их полностью. Методом K-средних расклассифицировать наблюдения в 2 класса, указать состав классов, проверить гипотезу существенности различий между центрами классов по каждой переменной, для самой значимой указать sl. Построить 3D график пронумерованных наблюдений, указать на нём полученные классы.

4. Разбить стандартизированные данные на 3 класса, проверить значимость.

Задание № 6.

5. Провести факторный анализ данных из задания № 1 на основе корреляционной матрицы, выделив методом главных компонент 2 фактора.
6. Найти нагрузки факторов на производительность труда (о1, о2).
7. Дать факторам названия и рассчитать коэффициенты информативности (о3, о4) на основе 2-х переменных.
8. Найти общность энерговооружённости и характерность уровня травматизма (о5, о6).
9. Указать оценки факторов для пятого предприятия (о7, о8).
10. Форма ответа: о1, о2, о3, о4, о5, о6, о7, о8.

Литература

Основная

1. Дубров А.М. Многомерные статистические методы для экономистов и менеджеров: Учебник для студентов вузов / А.М. Дубров, В.С. Мхитарян, Л.И. Трошин. М. : Финансы и статистика, 1998.

2. Сошникова Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер. М. : ЮНИТИ-ДАНА, 1999.

3. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. М. : ЮНИТИ, 2001.

4. Пугачев В.С. Теория вероятностей и математическая статистика / В.С. Пугачев. М. : ФИЗМАТЛИТ, 2002.

5. Большев Л.Н. Таблицы математической статистики / Л.Н. Большев, Н.В. Смирнов. М. : Наука, 1983.

Дополнительная

6. Тюрин Ю.Н. Статистический анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. М. : Инфра – М, 1998.

7. Руководство пользователя пакета программ SPSS v. 7.

8. Руководство пользователя пакета программ Statistica v. 6.

9. Справочник по прикладной статистике: в 2 т. / Под ред. Э. Ллойда, У. Ледермана. М. : Финансы и статистика, 1990.

10. Айвазян С.А. Прикладная статистика: Классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. М. : Финансы и статистика, 1989.

11. Хили Дж. Статистика. Социологические и маркетинговые исследования / Дж. Хили. СПб. : Питер, 2005.

12. Бородачёв С. М. Эконометрика : учебное пособие / Екатеринбург : УрФУ, 2011.

13. Бородачёв С. М. Методы математической статистики : учебное пособие / Екатеринбург : УрФУ, 2012.

Учебное издание

Бородачев Сергей Михайлович

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ

Редактор *Т. Н. Газитарова*

Подписано в печать 30.09.2009. Формат 60 х 84 1/16.
Бумага типографская. Плоская печать. Усл. печ. л. 4,29.
Уч.-изд. л. 3,0. Тираж 50 экз. Заказ

Редакционно-издательский отдел УГТУ – УПИ
620002, Екатеринбург, ул. Мира, 19
rio@mail.ustu.ru

Ризография НИЧ УГТУ – УПИ
620002, Екатеринбург, ул. Мира, 19